## THESIS TITLE

# Reduction of Latency in LTE Using Additional Control Region for Downlink Transmission and Prior Generation of Packets for Uplink Transmission

By


**F.M. Faijus Salehin Rifat (132464)**
**MD. Rumman Rafi (132461)**
**Mohammad Faisal Rahman Udoy (132483)**
**Naseef Zaki (132489)**


A Thesis Submitted to the Academic Facultyin Partial Fulfillment of the Requirements for the Degree of

**BACHELOR OF SCIENCE IN ELECTRICAL AND ELECTRONIC ENGINEERING**


Department of Electrical and Electronic Engineering

**Islamic University of Technology (IUT)**
Gazipur, Bangladesh


November 2017

# Reduction of Latency in LTE Using Additional Control Region for Downlink Transmission and Prior Generation of Packets for Uplink Transmission

Approved by:

--------------------------------------------

**Dr. Mohammad Tawhid Kawser**
Supervisor and Assistant Professor,
Department of Electrical and Electronic Engineering,
Islamic University of Technology (IUT),
Boardbazar, Gazipur-1704.

Date: 08/11/2017

## Table of Contents

## List of Tables

## List of Figures

# List of Acronyms

2G SECOND-GENERATION MOBILE
3G THIRD-GENERATION MOBILE
4G FOURTH-GENERATION MOBILE
5GFIFTH-GENERATION MOBILE
LTELONG TERM EVOLUTION
ARQ Automatic repeat request
CDMACODE DIVISION MULTIPLE ACCESS
DCCH DEDICATED CONTROL CHANNEL
DRA DYNAMIC RESOURCE ALLOCATION
D2D DEVICE-TO-DEVICE
FDMA FREQUENCY DIVISION MULTIPLE ACCESS
FDD FREQUENCY DIVISION DUPLEXING
FEC forward error correction
M2M MACHINE TO MACHINE
MCSMODULATION AND CODING SCHEME
MAC MEDIA ACCESS CONTROL
HSUPA HIGH-SPEED UPLINK PACKET ACCESS
RAT RADIO-ACCESS TECHNOLOGY
RTT ROUND TRIP TIME
RAN RADIO ACCESS NETWORK
PDCCH PHYSICAL DEDICATED CONTROL CHANNEL
PUCCH PHYSICAL UPLINK CONTROL CHANNEL
PUSCH PHYSICAL UPLINK SHARED CHANNEL
QAM QUADRATURE AMPLITUDE MODULATION
TTI TRANSMISSION TIME INTERVAL
TDD TIME DIVISION DUPLEX
QPSK QUADRATURE PHASE SHIFT KEYING
UE USER EQUIPMENT
UMTS UNIVERSAL MOBILE TELECOMMUNICATIONS SYSTEM

# Acknowledgements

With the best complements of Almighty Allah and Dedicated to...

Our Father and Mother.

# Abstract

The latency of LTE networks needs to be reduced to a darned low level to meet the expectation of 5G, which is the doorstep for future wireless communication.5G networks are currently being deployed over the experimental zone which has the potential to be the doorstep for future telecommunication. The drawbacks of this technology include high latency (compared to the defined expectations) and reliability. In this book we will try to reduce the latency over two proposals. Where both the layers can allocate resources over MCS level and the prediction of packets assuming previous uplink transmission in which we will try to show ways to reduce the latency.

# Chapter 1

# Introduction

The capabilities of Long Term Evolution (LTE) are now progressing towards the vision for fifth generation (5G) wireless technology and looking forward to better transforming and serving the users. Wireless networks have not been designed and engineered earlier to achieve such low latency and high reliability simultaneously. The time from a user sending a piece of data to the time when the user gets a response is sometimes more important than the bits offered. Of course, the latency requirements are very stringent for some applications while many other applications are delay tolerant. The latency can be measured by the time it takes for a small IP packet to travel from the terminal through the network to the internet server and back. 5G radio access network (RAN) stretching far beyond the year 2018, is expected to enable a highly scalable service experience by allowing people and machines to enjoy a gigabits data rate and at the same time, latency reduction in the order of 1 ms or even less. The V2X communication requires a latency of maximum 100 ns for critical messages. The state of the art LTE network can achieve a block error rate of 1-10% when end to end latency is limited to a few milliseconds. Thus, various methods have been proposed in order to achieve low latency in LTE but additional methods are needed to meet the stupendously low latency requirement for some applications. In this paper, to reduce latency in packet transmission for those highly delay sensitive applications.

We propose two separate methods. One method can be used for downlink transmission and the other can be used for uplink transmission. For downlink transmission, we propose introduction of an additional control region, continuously or intermittently spread over the whole sub frame. For uplink transmission, we propose generation of packets beforehand based on predicted modulation and coding scheme (MCS), which can obviate the need for four timeslots delay of transmission after scheduling.

5G Radio Access Technology (RAT), stretching far beyond year 2020, is expected to enable a highly scalable service experience by allowing people and machines to enjoy a virtually zero latency gigabits data rate on demand. 5G is targeting much higher throughput, spanning towards higher carrier frequencies and wider bandwidths, at the same time reducing energy consumption and costs compared to the existing 4G technologies. Latency reduction, such as physical round trip time (RTT) in the order of <1ms, can be seen as the main physical layer target to reach the high data rate requirements with feasible cost.

Densification of access points (or base stations), is the classical solution for providing increased capacity and improved coverage in wireless networks. This further leads to very dense small cell deployments and focuses the 5G air interface scope especially on local area (LA) networks. Consequently, multi-hop relaying in the form of wireless backhauling links, access links and device-to-device (D2D) links may be required in order to provide sufficient coverage for the multi-Gabs data flow. Time division duplexing (TDD) is considered a more attractive duplexing method than frequency division duplexing (FDD) in 5G dense deployment environment, as support burst traffic demand and new connection types, such as wireless backhaul or D2D. The rest of the paper focuses on TDD mode.

The air interface latency, concerning both control and data information, becomes critical in order to meet the latency requirements of <1 ms. This is especially important in a scheduled TDD system where several TDD cycles are required for delivering one round trip transmission related to control signaling or data transmission. Also, the large increase of 5G data throughput leads to the need of transmitting and processing a larger amount of data, consequently imposing demands on baseband processing. A baseband system can cope with the increased throughput demand by decreasing the latency. From the air interface perspective this essentially leads to the transmission of shorter blocks of data in time and wider blocks in frequency.

Besides the achievement of high data rates, latency reductions at the air interface level also become vital to enable energy savings and long battery life time. Fast transitions between sleep and active modes, short active time with high data rate together with low sleep mode power consumption are required to guarantee multiple years of lifetime for a small low cost battery. On air interface level, the need to transmit control and user data quickly in time domain leads again to the demand of fast link direction switching and to short transmission time interval (TTI) length.

The air interface latency of TDD Long Term Evolution Advanced (LTE-A) [2][3] is limited by its physical frame structure. It is possible to include at maximum two uplink (UL) / downlink (DL) switching points inside one 10 ms. radio frame, which sets the hard limit for the air interface latency. This is clearly not achieving the 5G physical layer latency target. Evolutions of LTE-A will not be able to support major latency reductions due to the restrictions of incremental evolution. For example, changes in the numerology and frame structure design for enabling latency reductions can hardly be introduced for backwards compatibility reasons. Consequently, there is a need of a new 5G air interface enabling the required physical layer latencies.

Current generation of wireless technologies is being evolved toward a fifth generation (5G) for better serving end users and transforming our society. In addition to enhancing now essential services such as voice and mobile broadband, there has been strong momentum for improving wireless technologies for machine-type (or machine-to-machine) communications. A shared vision is that by providing an Internet of Things (IoT) to connect various types of machines, devices, or sensors, the productivity and efficiency in industries, and our society in general, will be improved.

A range of 5G machine-type communication (MTC) use cases, such as smart grid power distribution automation, industrial manufacturing and control, intelligent transportation systems, remote control of machines, and remote surgery, are characterized by the need for reliable real-time communication with high requirements on latency, reliability, and availability. For example, the very stringent requirements for factory automation include reliability corresponding to a maximum block error rate of 10-9 and latency down to a millisecond (ms.) or less.

Traditionally, wireless networks have not been designed and engineered to achieve such low latency and high reliability simultaneously. For example, the state-of-the-art LTE network can achieve a block error rate of approximately 1-10% when the end-to-end latency is limited to a few milliseconds. For relaxed latency requirements, LTE can eventually achieve any level of reliability through retransmissions at various protocol layers. However, achieving high reliability in a fading radio channel under tight latency constraints presents a new challenge and new radio access designs need to be developed.

To explore the viability of using radio technology for reliable real-time communication and to outline important design choices of such a solution, we will investigate communication for automation of manufacturing cells, a scenario with some of the toughest requirements on latency and reliability. Factory automation is real-time control of machines and systems in fast production. Measurements and control commands need to be applied quickly, and fail-safe transfer of sensor and actuator signals is important. Small amounts of data, 10-20 bytes, are communicated with latency as low as 0.5-1 ms. and reliability of not higher than 10-9 probability of failed packet delivery within the latency bound.

To enable an end-to-end latency of 1 ms. or less, the entire system needs to be designed for low latency, including multiple access technique, radio access network, backhaul, storage, etc. In this work, we assume that the transmission time needs to be on the order of 100 μs to allow for processing delays in the transmitter, receiver, scheduler, and other system components. Therefore we target a 100 μs transmission time, and this time interval must also include any retransmissions of the same packet. The reliability requirement targeted will be 10-9 residual block error rate after potential.

# Chapter 2

# Overview of latency

## *2.1 Introduction:*

The time from a user sending a piece of data, request download or a webpage to load, to the time when the user gets a response, which is sometimes more important than the bit rate offered. The latency can be measured by the time it takes for a small IP packet to travel from the terminal through the network to the internet server, and back. That measure is called round trip time and is illustrated in Figure:

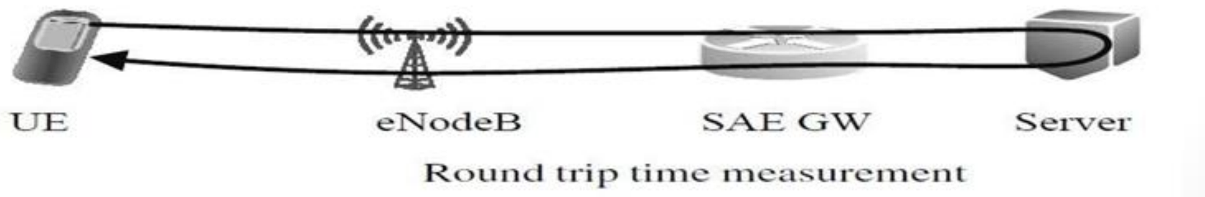


Figure 2-2-1 Round trip latency

### 2.1.1 Why does latency matter?

System latency is more important than the actual peak data rates for many IP based applications.

User are more dependent on applications which is reliable on low latency such as:

1- Games
2-VOIP
3-Mail & file Sync
4-Application sharing
5-Video & voice conference over IP

### 2.1.3 Latency in numbers

Here is a table to add up

Table 2. 1 latency in numbers

| | **Application** | **Degree of symmetry** | **Data rate** | **E2e one way delay** | **Delay variation within a call** | **Information loss** |
|---|---|---|---|---|---|---|
| **Audio** | Conversational voice | Two-way | 4-25 kb/s | <150 msec preferred<br><400 msec limit | < 1 msec | < 3% FER |
| **Video** | Videophone | Two way | 32-384 kb/s | <150 msec<br><400 msec limit<br>Lip synch<100 msec | 2.1.4 | < 1% FER |
| **Data** | Telemetry | Two way | <28.8 kb/s | <250 msec | N.A | Zero |
| **Data** | Real-time games | Two way | <60 kb/s | <75 msec preferred | N.A | <3% FER<br><5% FER limit |
| **Data** | Telnet | Two-way (Asymmetric) | <1 KB | <250 msec | N.A | Zero |

## *2.2 RTT in different technology*

GSM/EDGE 150ms
HSPA 100ms
HSPA+ 50ms
LTE 20ms
Target is to have it as low as 5ms

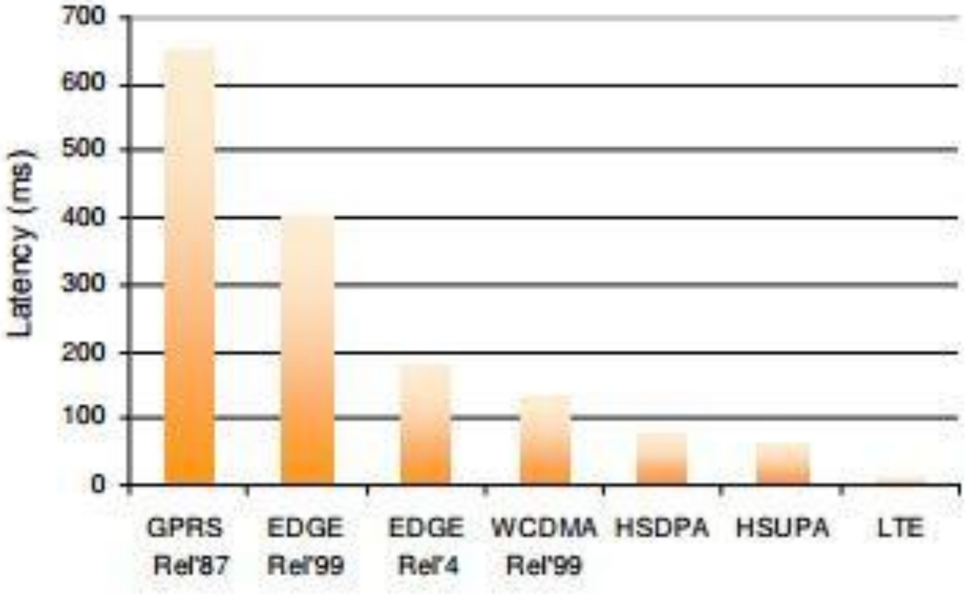


Figure 2-2-2 Latency Vs Generation

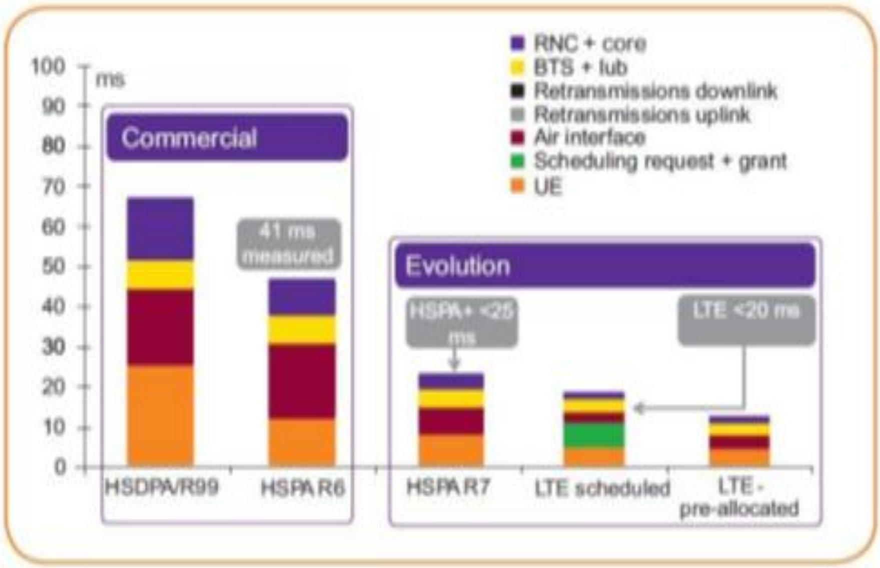


Figure 2-2-3 Latency in HSPA LTE

## *2.3 Latency and throughput*

Latency and throughput are two critical performance metrics of a communication network. Recently, a lot of attention has been focused on improving throughput (or spectral efficiency) of Wireless Wide Area Networks (WWANs) through the use of physical and MAC layer techniques, such as higher order modulation, MIMO and aggregation of bandwidth (Multi-carrier). While some data applications directly benefit from the higher data rates, for many applications high data rates do not translate to improved user experience unless the latency is low. In this paper, we examine the Control plane (C-plane) and User plane (U-plane) latencies in an HSPA data system. We show that significant C-plane latency reduction can be achieved in HSPA by carrying signaling on HS-DSCH and E-DCH channels (as opposed to dedicated channels, which is the practice now). We also compare the C-plane and U-Plane latencies of HSPA and LTE, which have comparable spectral efficiency.

3rd Generation Partnership Project (3GPP) has standardized Code Division Multiple Access (CDMA) based packet-switched air-interfaces for downlink and uplink, called High-Speed Downlink Packet Access (HSDPA) and High-Speed Uplink Packet Access (HSUPA) respectively, together referred to as High Speed Packet Access (HSPA) [2][4]. In today's HSPA deployments, signaling messages are typically carried on dedicated channels. These dedicated channels are typically low rate channels (3.4 kbps), which allows larger spreading factors to be used so that more users can share the cell's resources. Since some signaling messages tend to be fairly large, the use of these low rate dedicated channels leads to call setup latencies that are in the range of a few seconds. Newer HSPA deployments are considering carrying signaling on HSPA channels (i.e., HS-DSCH on the downlink and E-DCH channels on the uplink). Such configurations not only allow better statistical use of power and code resources, but also allow signaling to benefit from higher rate HSPA channels. We will show that significant improvements to call setup latency can be achieved by carrying signaling on HSPA channels.

In parallel with the evolution of HSPA, 3GPP has also standardized an Orthogonal Frequency Division Multiplexing (OFDM) based air-interface, called Long Term Evolution (LTE) [3]. Though the spectral efficiencies of the two systems are comparable, Rel-9 LTE provides higher peak rates (due to its wider 20MHz bandwidth) compared to Rel-8/Rel-9 HSPA, which allows a User Equipment (UE) to receive on 10 MHz bandwidth with carrier aggregation. This paper focuses on the call setup and user plane latencies ofthe two systems. Call setup latency determines how quickly the user can start to receive service, while user plane latency is important since the performance of applications such as web browsing is very sensitive to the Round Trip Time (RTT) to the server.

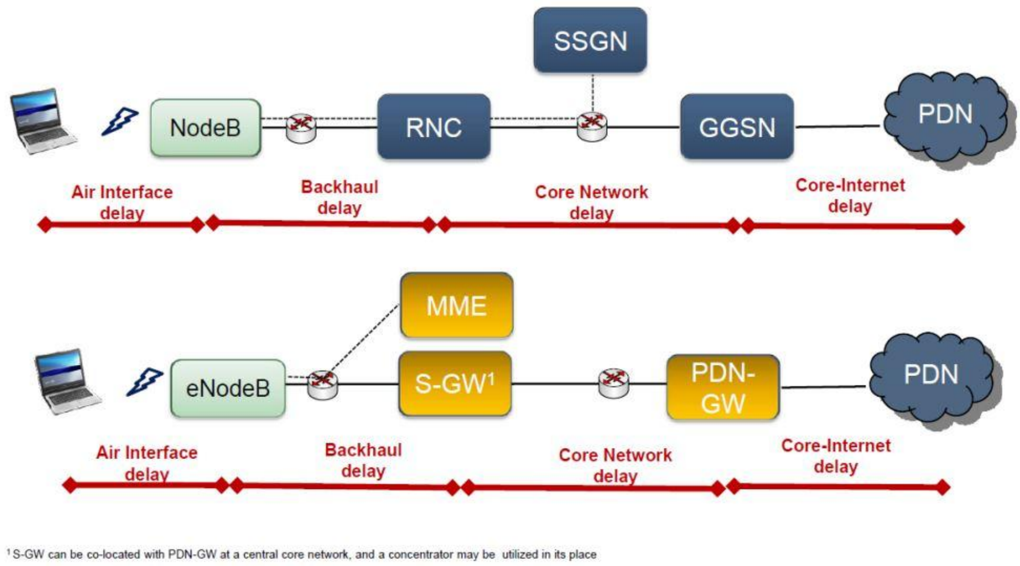


Figure 2-4Transport Network Architecture for HSPA and LTE

## *2.4 System ArchitectureEvolution (SAE)*

The LTE was designed with less NE, by integrating the Radio Network controller functionality in Node/eNodeB it enables data traffic to by-pass the RNC and SGSN and network elements to connection Radio Access Network (RAN) to the core packet network directly. The RTT will improve because there is no Iub related transport set-up delay. The transport connection is on all the time and the IP packets are sent immediately.

### 2.4.1 User Plane Latency:

The latency can be measured by the time it takes for a small IP packet to travel from the terminal through the network to the internet server, and back. That measure is called round trip time. (LTE for UMTS) U-Plane latency is defined as one-way transmit time between a packet being available at the IP layer in the UE/E-UTRAN (Evolved UMTS Terrestrial Radio Access Network) edge node and the availability of this packet at the IP layer in the EUTRAN/UE node. U-Plane latency is relevant for the performance of many applications.

Loosely speaking, User plane (U-plane) latency is the delay seen by an application in exchanging data with a server. A metric that is commonly used to characterize U-plane latency is the PING delay, i.e., the delay from a UE sending a small PING to the first IP node in the network and receiving the PING response.

This PING latency is the definition of U-plane latency used in this paper. Control plane latency: Control plane (C-plane) latency, also known as call setup latency, is the latency for a User Equipment (UE) to transition to a state where it can send/receive data. The definition of call setup latency merits further clarification. The 3GPP specification for HSPA allows the UE to camp1 with low battery consumption in either RRC idle or CELL_PCH (or URA_PCH, which is similar to CELL_PCH for the purposes of this paper) states. In these states, the UE only needs to listen to the network in pre-assigned "on" periods, while the UE can save battery by shutting off its receiver during the rest of the DRX (Discontinuous Reception) cycle.

Camping in CELL_PCH state has a number of benefits: the signaling cost and delay to transition the UE from CELL_PCH to a state where the UE can send/receive data (i.e., CELL_FACH) is very small compared to starting from idle state, while the UE's battery consumption in CELL_PCH and idle states is similar (assuming same DRX cycles for both). Given the two types of camped states (idle and CELL_PCH) allowed by the HSPA specification, we consider two definitions of call setup, one from RRC idle to CELL_DCH (we refer to this as call setup from idle) and one from CELL_PCH to CELL_FACH (we refer to this as call setup from connected).

### 2.4.2 Control Plane Latency:

Control plane deals with signaling and control functions. C-Plane latency is measured as the time required for the UE (User Equipment) to transit from idle state to active state. In idle state, the UE does not have an RRC connection. Once the RRC is setup, the UE transitions to connected state and then to the active state when it enters the dedicated mode. Enhanced FACH & RACH. The expected packet call setup time with Release 7 will be below 1s. Once the packet call has been established, user data can flow on HSDPA/HSUPA in Cell DCH state. The idea in Release 7 Enhanced FACH and Release 8 Enhanced RACH is to use the Release 5 and Release 6 HSPA transport and physical channels also in the Cell_ state for improving the end user performance and system efficiency.

Legacy HSPA networks typically transitioned the UE to RRC idle state once data transfer was finished, so UEs experienced the latency of call setup from idle. However, with the deployment of CELL_PCH, we expect that UEs will typically camp in CELL_PCH, and only very long continuous inactivity (of the order of tens of minutes or more) may cause them to be transitioned to RRC idle. Thus, in future HSPA networks, we expect CELL_PCH (as opposed to idle) to be the default camped state from which the UE will initiate most of its data transactions. While CELL_PCH is supported in the specifications since R99, an Enhanced version of CELL_PCH was added in Rel 7. Enhanced CELL_PCH allows the UE to retain a dedicated H-RNTI in CELL_PCH state, which means that the UE is primed to send/receive data. In the case of LTE, the only two RRC states are idle and connected. In connected state, the UE is allowed to camp with a long discontinuous receive (DRX) cycle: for the purposes of this paper, this can be considered the equivalent of the CELL_PCH state in HSPA. Thus, for LTE, the term call setup from idle refers to a transition from RRC idle to a state where data can be sent/received, and the term call setup from connected refers to a transition from RRC connected with a long DRX cycle to a state where data can be sent/received.

### 2.4.3 Resource Allocation

The first one, two or three symbols out of fourteen symbols of every sub frame are used as control region in LTE and so, the control region is 0.0719 ms. 0.1432 ms. or 0.2146 ms long, respectively, assuming normal cyclic prefix. The channels which carry data exist in the rest part of 1 ms long sub frame and they are physical downlink shared channel (PDSCH) in downlink and physical uplink shared channel (PUSCH) in uplink. The physical downlink control channel (PDCCH), located in the control region, allocates resources for PDSCH and PUSCH and specifies MCS for the data therein. The dynamic scheduling, suitable for bursty payloads, allocates uplink and downlink resources subframe by subframe. The CRC of a PDCCH instance is scrambled by the C-RNTI of the UE, which is to be allocated resources. Since the control region is located right at the beginning of a subframe, if any packets arrive after the control region is complete, they cannot be scheduled until the next subframe. This can cause significant delay.

In the case of uplink, if PDCCH signaling is sent on subframe n, then it allocates resources on PUSCH on the subframe n+4. This time interval allows the UE to dequeue its data and generate its transport block according to the specified MCS. The UE sets up an appropriate timing offset between its transmitted radio frames in uplink and its received downlink radio frames as shown in Fig. 2.3-1. The eNodeB sends a timing advance command to the UE with necessary adjustment of timing offset. The timing offset between uplink and downlink radio frames can range from 0 to 0.67 ms with a granularity of 0.52 µs. The

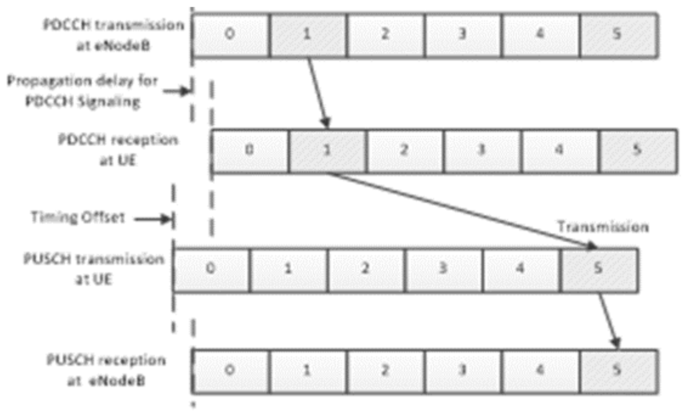


Figure 2-5 Timing offset for uplink transmission

maximum possible timing offset, 0.67 ms, supports 100 km cell size.

# Chapter 3

# LITERATURE REVIEW

## *3.1 Overview*

The capabilities of Long Term Evolution (LTE) are now progressing towards the vision for fifth generation (5G) wireless technology and looking forward to better transforming and serving the users. Wireless networks have not been designed and engineered earlier to achieve such low latency and high reliability simultaneously.

There has been numerous work going on to achieve low latency whether by changing the network architecture or by engaging new methods to minimize delay over packet transmission.

## *3.2 Overview of the subject*

As it has been already been mentioned in earlier sections that the main objective of this research is to reduction in 5G for both uplink and downlink communication. So far this has been a major concern for the telecommunication sector as it will open up totally a new dimension in the advanced technology field.

The expertise & industries are working on areas like renovating the control plane and
Modifying the transmission time interval (TTI) length to achieve low latency. In the upcoming sections a brief discussion will be given of the proposed solution for the low latency and the merits and demerits & how this paper has tried to overcome the shortcomings of those research that has been made.

## *3.3 Categorization of sources*

For LTE, latency reduction techniques were studied. Based on the study outcome, a layer 2 solution was specified in Rel. 14 [6], and a layer 1 solution is to be specified. The layer 2 solution is to allow the user equipment (UE) to skip uplink transmission if the UE has no data. More specifically,a network can configure/schedule uplink resources for a UE without taking into account data buffer of the UE, and then the UE can decide whether to transmit or skip uplink data depending on whether data is available in the UE buffer. In the legacy LTE, the UE shall send data in response to an allocated UL dynamic or configured grant even if no data is available in the UE buffer. Allowing uplink transmission skipping decreases uplink interference and improves UE energy efficiency, and makes semi-persistent resource allocation more realistic.

The layer 1 solution includes two sub-solutions:
1. Shortened processing time for 1 ms TTI
2. Shortened TTI with shortened processing time

In general, shorter TTI length enables shorter processing time. Therefore, 2 is expected to realize further latency reduction compared to 1.

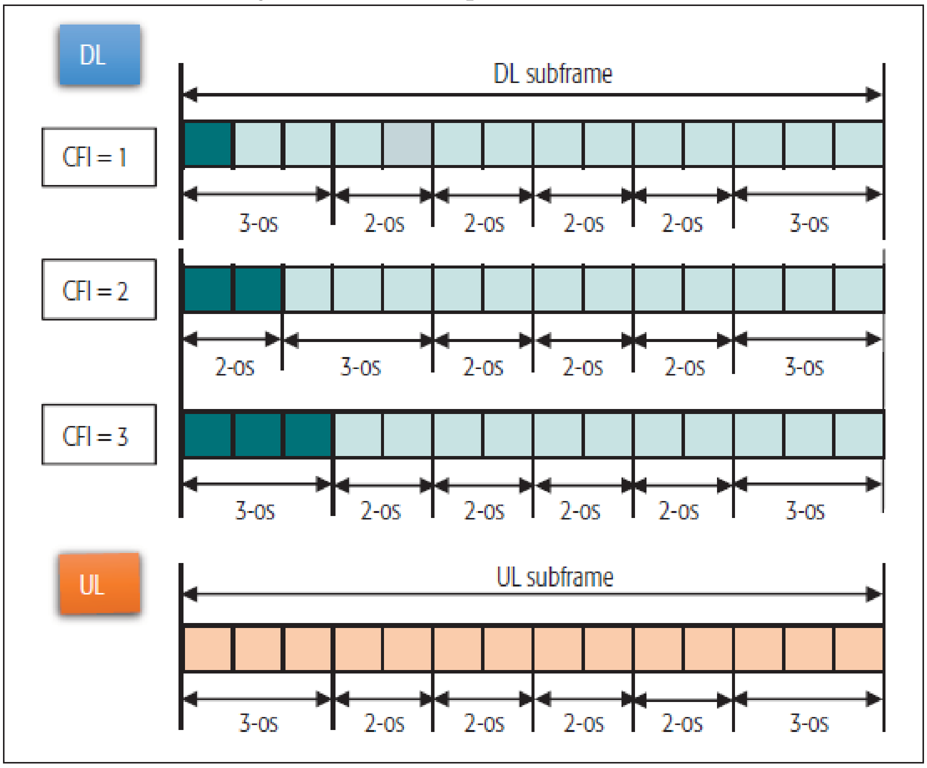


Figure 3-1 DL and UL short-TTI patterns for 2-symbol (2-os) sTTI.

For TTI shortening in NR, there are two approaches. The first approach is to allow data TTI having smaller number of OFDM symbols (e.g., 1 or 2), similar to LTE short-TTI. The difference from LTE short-TTI is that backward compatibility is not required for NR [10]. Therefore, it is possible to design a whole NR system, including control/data channels, reference signals, and related UE behaviors, from the beginning, such that various TTI lengths can flexibly be applicable depending on the service type of each UE. UEs with various TTI lengths shall be able to coexist in the same carrier efficiently. In order to realize such flexible TTI durations in a unified manner, reference signals for data demodulation should be confined within a limited number of OFDM symbols (e.g., 1 symbol). As long as the reference signals for demodulation is available, the number of OFDM symbols in which a data spans can be shorter (Fig. 2). Then, different TTIs having different numbers of OFDM symbols can be multiplexed on the same carrier in a flexible manner as illustrated in Fig. 3. This design principle ensures forward compatibility; that is, future new services requiring a certain data rate, latency, reliability, and so on can be realized by the framework.

Together with the flexible TTI duration, for NR, a new concept is now under consideration: sub-blocking of a TTI. More specifically, the transport block for a given TTI is divided into sub-blocks, and each sub-block is encoded and modulated independently.

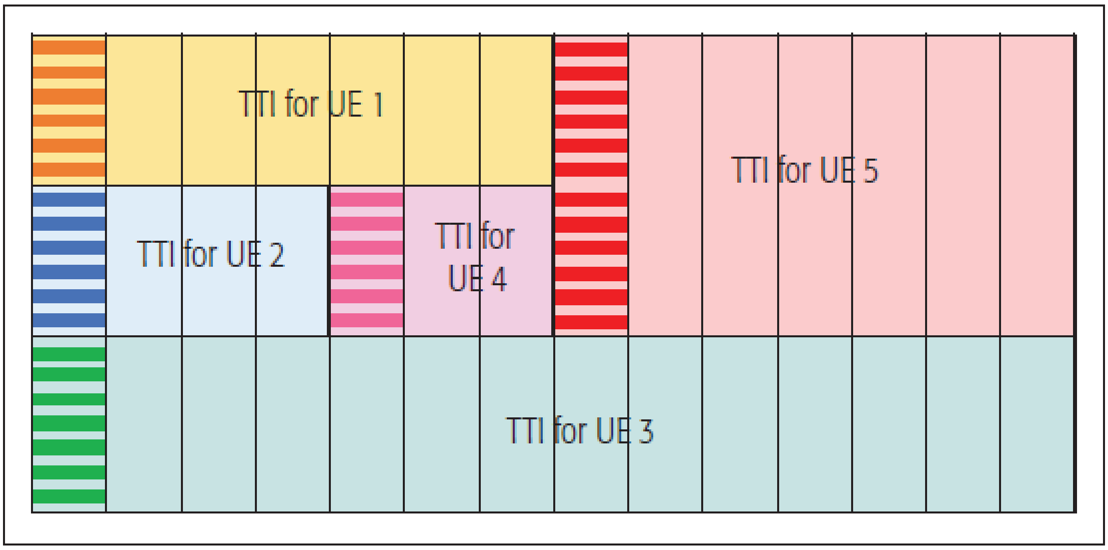


Figure 3-2 Multiplexing data channels having different TTI durations on one carrier.

In this section, an overall system structure to provide short-TTI functionality in an LTE system is introduced.

First of all, we consider 1, 2, 3/4, and 7 OFDM symbols as the length of short TTI, where the OFDM symbol duration of LTE systems is about 71.5 us. In order not to make no interference to legacy LTE UEs, the eNB should not allocate any data for short-TTI UEs in those RE used for PDCCH and CRS in the legacy LTE system. Since PDCCH of the legacy LTE system exists system bandwidth, a short TTI has to be out of the legacy PDCCH region to preserve backward compatibility. The frequency-time resources for short-TTI PDCCH (sPDCCH) and short- TTI physical downlink shared channel (sPDSCH) are shown in figure.

In follows, we provide the preliminary system-level evaluation results based on the proposed short-TTI structures in the previous subsection. For control overhead assumption, we consider 4 CCE (control channel element) s and 8 CCEs for 1-symbol TTI and 2-symbol TTI, respectively, where 1 CCE equals to 36 resource elements in LTE systems. For 3/4- symbol TTI, the control of the first TTI is transmitted in the legacy PDCCH while the other TTIs are assumed to have 12 CCEs in each sTTI. For 7-symbol TTI, i.e., slot TTI, the second slot has 12 CCEs for sPDCCH. The baseline with subframe TTI assumes that the control channel takes only 2-symbols in the PDCCH region. If we consider uplink (UL) scheduling as well, the control overhead should be increased. However, the UL scheduling control overhead is not considered for simplicity. For system-level evaluations, we assume to use 7 macro sites, 3 sectors per site, 2 GHz carrier frequency, 10 MHz system bandwidth, 500 m interstice distance, 46 dBm total base station transmission power, ITU UMa path loss and shadowing model, and MMSE-IRC receiver. Furthermore, we assume that HARQ RTT and UL access delay are reduced proportionally to TTI length and core network delay is 6 ms.

Then, we observe the user perceived throughput that is defined as file size divided by the time needed to download the file starts when the packet is generated, where the evaluation results are shown in Tables 1 and 2. As can be seen, the evaluation with the smaller file sizes shows much performance gain in mean UPT and mean latency for short TTI. Also, even cell-edge UEs with short TTI can have much UPT gain for 100 kbits file size. For large-sized files, cell-edge UEs have negative gains in UPT performance. As RU is gettinghigher, performance gain becomes bigger for a 100 kbit-sized file while performance gain becomes smaller for the other file size. As can be seen in, after studying several techniques, LTE Release-15 adopts 2 symbols and 7 symbols TTI lengths. Regarding the effect of resource utilization (RU), when TCP is not considered, if RU increases, thelatency could normally increase so that UPT could decrease. However, in this evaluation with TCP model, it turns out that the performance difference for various RU is not as much as the case without TCP modelling. Furthermore, as RU is getting higher, performance gain becomes bigger for a 100 kbit-sized file while performance gain becomes smaller for otherfile sizes. This is because a file is distributed over longer period by TCP slow start.

## *3.4 Arguments and discussions*

The proposals that have been discussed in the previous section and it has some similarities with our proposed solution. Though we are aligned with the technique of shortening TTI , but this paper has uniquely addressing the provision of additional control region for downlink transmission.

In this paper, the present control region, restricted to the beginning of a subframe, is termed as primary control region. We propose that in addition to the primary control region, a secondary control region will be used throughout the rest part of the subframe. The secondary control region takes up symbols that can range from the symbol next to the primary control region up to the symbol before the last. The secondary control region will be used for quick allocation of downlink resources. Thus, the primary control region does not allocate for all resource blocks (RBs). It rather leaves RBs out in the time-frequency grid of the subframe and these RBs will be considered broken. A broken RB will contain a short resource blocks (SRB) in every symbol period as shown in Fig. 2. Thus, a SRB will have 12 resource elements (REs).

The secondary control region will allocate SRBs symbol by symbol. Like the primary control region, the secondary control region will also be comprised of strongly coded PDCCH instances and the CRC of a PDCCH instance will be scrambled by the C-RNTI of the UE. If any downlink packets for highly delay sensitive applications arrive after the primary control region is complete, the next available symbol in the secondary control region will allocate resources for them and preferably, it will allocate SRBs in its next symbol minimizing the delay. However, if the secondary control region does not find sufficient packets from highly delay sensitive applications for allocation of all SRBs in its next symbol, it can allocated from less delay sensitive applications or even from delay tolerant applications and thus, avoid wastage of SRBs.

The secondary control region itself will be allocated on SRBs and the allocation will be specified by a PDCCH instance in the primary control region. The secondary control region can be either continuously or intermittently spread over the whole subframe. The secondary control region should be allocated on frequencies that are suffering from less interference and noise. A broken RB can be shared, on different SRBs, by both the secondary control region and the downlink data allocation.

The user equipment (UE) may have its data available for transmission and it cannot generate packets until it receives PDCCH signaling because the specified MCS is unknown. As shown in Fig. 1, because of the timing offset, when the UE receives PDCCH signaling on subframe n, the subframe n on PUSCH can already elapse some good extent of time, albeit less than the length of one subframe. This is because the maximum possible timing offset is 0.67 ms. Thus, the next available subframe on PUSCH is n+1. We propose that the UE generates uplink packets based on predicted MCS before the reception of PDCCH signaling. If PDCCH signaling is sent on subframe n, then it allocates resources on PUSCH on the subframe n+1. This can reduce the latency by 3 ms

This two proposals further improves the TTI as there is no data loss or buffer time as the resource can be allocated just after the subframe it has been allocated to which will further discuss on chapter 5.

# Chapter 4

# LTE

## *4.1 Introduction:*

In telecommunication, Long-Term Evolution (LTE) is a standard for high-speed wireless communication for mobile devices and data terminals, based on the GSM/EDGE and UMTS/HSPA technologies. It increases the capacity and speed using a different radio interface together with core network improvements. The standard is developed by the 3GPP (3rd Generation Partnership Project) and is specified in its Release 8 document series, with minor enhancements described in Release 9. LTE is the upgrade path for carriers with both GSM/UMTS networks and CDMA2000 networks. The different LTE frequencies and bands used in different countries mean that only multi-band phones are able to use LTE in all countries where it is supported.

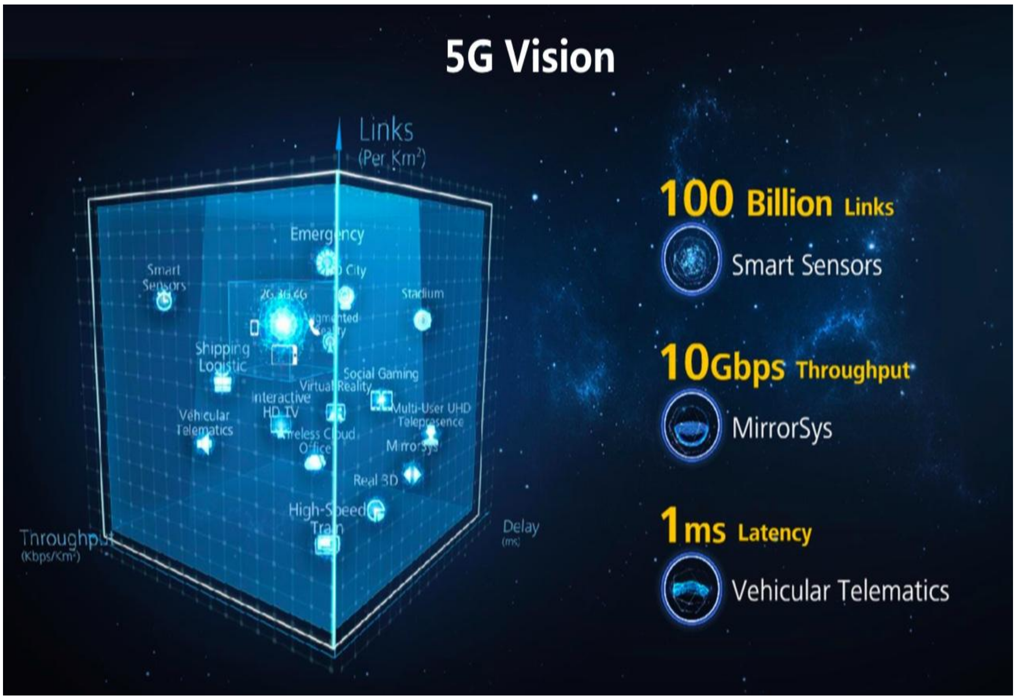


Figure 4-1 5G vision

LTE is commonly marketed as 4G LTE, but it does not meet the technical criteria of a 4G wireless service, as specified in the 3GPP Release 8 and 9 document series, for LTE Advanced. The requirements were originally set forth by the ITU-R organization in the IMT Advanced specification. However, due to marketing pressures and the significant advancements that WiMAX, Evolved High Speed Packet Access and LTE bring to the original 3G technologies, ITU later decided that LTE together with the aforementioned technologies can be called 4G technologies. The LTE Advanced standard formally satisfies the ITU-R requirements to be considered IMT-Advanced.To differentiate LTE Advanced and WiMAX-Advanced from current 4G technologies, ITU has defined them as "True 4G".

Machine-to-machine communication has attracted a lot of interest in the mobile communication industry and is under standardization process in 3GPP. Of particular interest is LTE-Advanced support for various M2M service requirements and efficient management and handling of a huge number of machines as mobile subscribers. In addition to the higher throughput, one of the main advantages of LTE/LTE-A in comparison with the previous cellular networks is the reduced transmission latency, which makes this type of networks very attractive for real-time mobile M2M communication scenarios. This paper presents a M2M system architecture based on LTE/LTE-A and highlights the delays associated with each part of the system. Three real-time M2M applications are analyzed and the main latency bottlenecks are identified. Proposals on how the latency can be further reduced.

LTE Release 8 is one of the primary broadband technologies based on OFDM, which is currently being commercialized. LTE Release 8, which is mainly deployed in a macro/microcell layout, provides improved system capacity and coverage, high peak data rates, low latency, reduced operating costs, multi-antenna support, flexible bandwidth operation and seamless integration with existing systems. LTE-Advanced (also known as LTE Release 10) significantly enhances the existing LTE Release 8 and supports much higher peak rates, higher throughput and coverage, and lower latencies, resulting in a better user experience. Additionally, LTE Release 10 will support heterogeneous deployments where low-power nodes comprising picocells, femtocells, relays, remote radio heads, and so on are placed in a macrocell layout. The LTE-Advanced features enable one to meet or exceed IMT-Advanced requirements. It may also be noted that LTE Release 9 provides some minor enhancement to LTE Release 8 with respect to the air interface, and includes features like dual-layer beamforming and time-difference- of-arrival-based location techniques. In this article an overview of the techniques being considered for LTE Release 10 (aka LTEAdvanced) is discussed. This includes bandwidth extension via carrier aggregation to support deployment bandwidths up to 100 MHz, downlink spatial multiplexing including single-cell multi-user multiple-input multiple-output transmission and coordinated multi point transmission, uplink spatial multiplexing including extension to four-layer MIMO, and heterogeneous networks with emphasis on Type 1 and Type 2 relays. Finally, the performance of LTEAdvanced using IMT-A scenarios is presented and compared against IMT-A targets for full buffer and bursty traffic model.

The detailed understanding of packet delays in modern wireless networks is crucial to optimize applications and protocols. We conducted high precision latency measurements in operational LTE and HSPA networks, deploying a hybrid approach of active probing and withe-box testing. It allowed us to separately assess the one-way delay contributions of the radio access network and the core network for both technologies. The results show that LTE outperforms HSPA in the case of medium to high data rates. However, due to differences in the radio access procedures, the HSPA uplink connection offers lower delay for specific traffic patterns. A comparison between our measurement results and the requirements for delay sensitive applications exhibits that LTE is not (yet) the generally preferable technology. Hence, further optimizations of the LTE scheduling and resource allocation policies are required to fully exhaust all feasible latency improvements.

Goals for the evolved system include support for improved system capacity and coverage, high peak data rates, low latency, reduced operating costs, multi-antenna support, flexible bandwidth operations and seamless integration with existing systems. To reach these goals, a new design for the air interface including control channel is envisioned. This paper provides a preliminary look at an efficient downlink control channel design to reduce the overhead required to support data transmission. Initial performance results show that close to optimal system performance can be achieved with downlink control overhead of less than 14%.

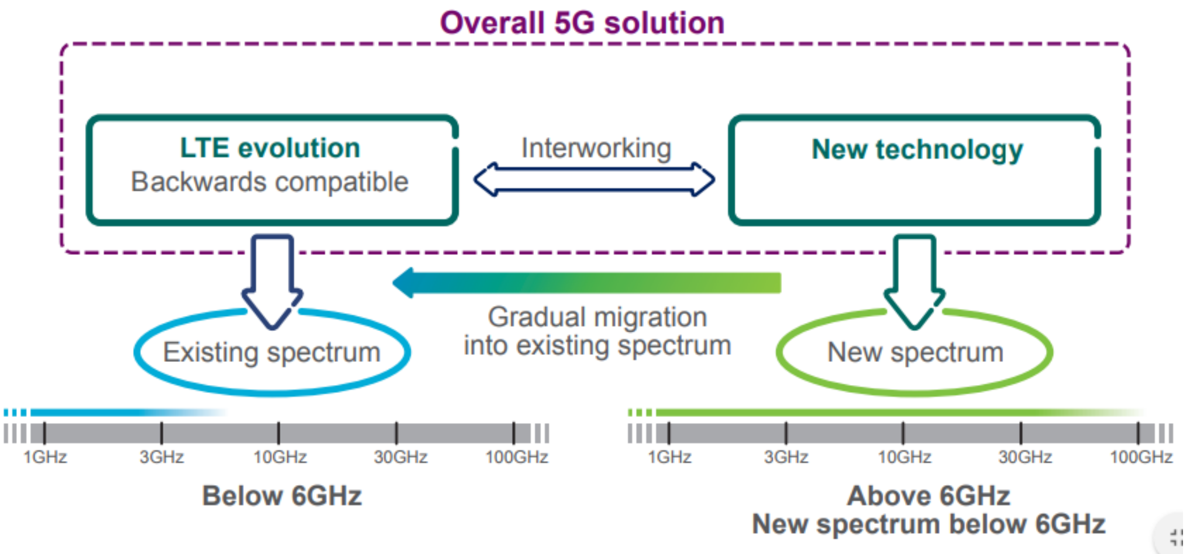


Figure 4-2Overall 5G wireless-access solution of LTE evolution and new technology

## 4.2 5G – requirements and capabilities

In order to enable connectivity for a very wide range of applications with new characteristics and requirements, the capabilities of 5G wireless access must extend far beyond those of previous generations of mobile communication. These capabilities will include massive system capacity, very high data rates everywhere, very low latency, ultra-high reliability and availability, very low device cost and energy consumption, and energy-efficient networks.

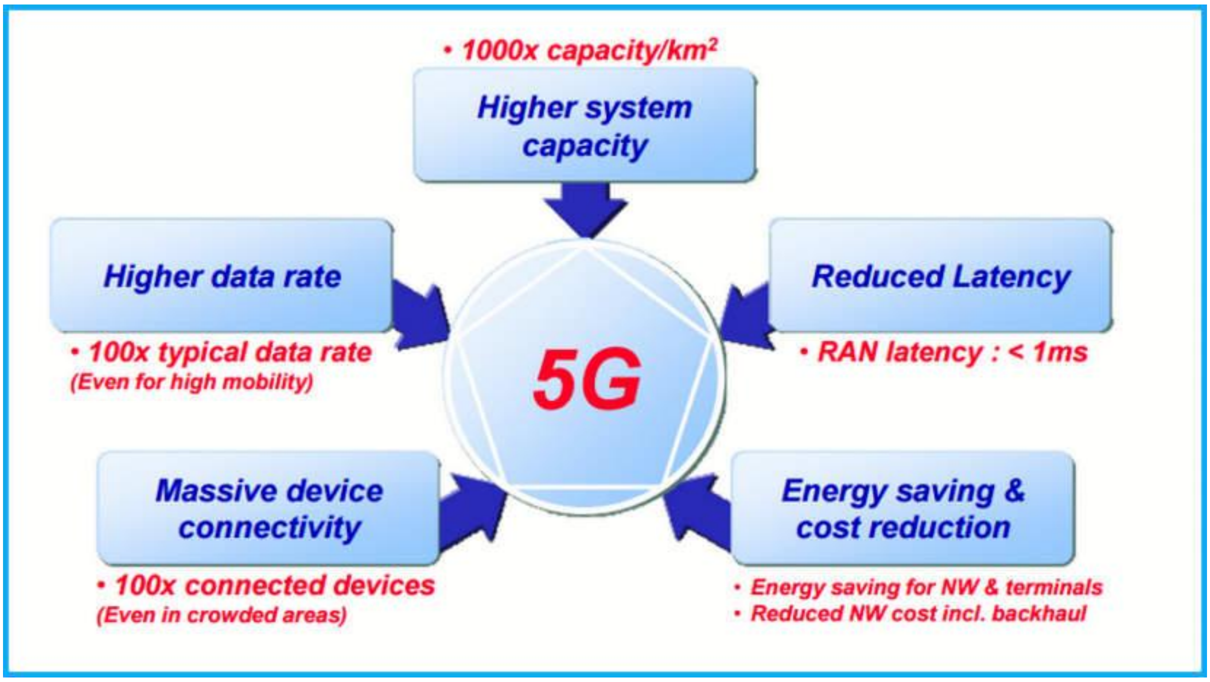


Figure 4-3 5G requirements

## *Massive system capacity*

Traffic demands for mobile-communication systems are predicted to increase dramatically. To support this traffic in an affordable way, 5G networks must deliver data with much lower cost per bit compared with the networks of today. Furthermore, the increase in data consumption will result in an increased energy footprint from networks. 5G must therefore consume significantly lower energy per delivered bit than current cellular networks. The exponential increase in connected devices, such as the deployment of billions of wirelessly connected sensors, actuators and similar devices for massive machine connectivity, will place demands on the network to support new paradigms in device and connectivity management that do not compromise security. Each device will generate or consume very small amounts of data, to the extent that they will individually, or even jointly, have limited impact on the overall traffic volume. However, the sheer number of connected devices seriously challenges the ability of the network to provision signaling and manage connections.

## *Very high data rate everywhere*

Every generation of mobile communication has been associated with higher data rates compared with the previous generation. In the past, much of the focus has been on the peak data rate that can be supported by a wireless-access technology under ideal conditions.

However, a more important capability is the data rate that can actually be provided under real-life conditions in different scenarios.

> 5G should support data rates exceeding 10Gbps in specific scenarios such as indoor and dense outdoor environments.
> Data rates of several 100Mbps should generally be achievable in urban and suburban environments.
> Data rates of at least 10Mbps should be accessible almost everywhere, including sparsely populated rural areas in both developed and developing countries.

# *Very low latency*

Very low latency will be driven by the need to support new applications. Some envisioned 5G use cases, such as traffic safety and control of critical infrastructure and industry processes, may require much lower latency compared with what is possible with the mobile-communication systems of today. To support such latency-critical applications, 5G should allow for an application end-to-end latency of 1ms or less, although application-level framing requirements and codec limitations for media may lead to higher latencies in practice. Many services will distribute computational capacity and storage close to the air interface. This will create new capabilities for real-time communication and will allow ultra-high service reliability in a variety of scenarios, ranging from entertainment to industrial process control.

## *Ultra high reliability and availability*

In addition to very low latency, 5G should also enable connectivity with ultra-high reliability and ultra-high availability. For critical services, such as control of critical infrastructure and traffic safety, connectivity with certain characteristics, such as a specific maximum latency, should not merely be ‘typically available.’ Rather, loss of connectivity and deviation from quality of service requirements must be extremely rare. For example, some industrial applications might need to guarantee successful packet delivery within 1 ms with a probability higher than 99.9999 percent.

## *Very low cost and energy consumption:*

Low-cost, low-energy mobile devices have been a key market requirement since the early days of mobile communication. However, to enable the vision of billions of wirelessly connected sensors, actuators and similar devices, a further step has to be taken in terms of device cost and energy consumption. It should be possible for 5G devices to be available at very low cost and with a battery life of several years without recharging.

## *Energy efficient network*

While device energy consumption has always been prioritized, energy efficiency on the network side has recently emerged as an additional KPI, for three main reasons:
> Energy efficiency is an important component in reducing operational cost, as well as a driver for better dimensioned nodes, leading to lower total cost of ownership.
> Energy efficiency enables off-grid network deployments that rely on medium-sized solar panels as power supplies, thereby enabling wireless connectivity to reach even the most remote areas.
> Energy efficiency is essential to realizing operators' ambition of providing wireless access in a sustainable and more resource-efficient way.
The importance of these factors will increase further in the 5G era, and energy efficiency will therefore be an important requirement in the design of 5G wireless access.

## *4.3 5G Vision:*

i) 100 Billion Connections,

ii) 1 ms Latency, and

iii) 10 Gbps Throughput

Mobile networks are changing the way people communicate and access information. Network access at anytime and anywhere is transforming the telecom industry. In the near future, wireless network access will eventually prevail. 5G technology will enable flexible, reliable, and secure wireless networks to connect people with all applications, services, and things, thus leading human race into the era of "Everything on Mobile".

In the "Everything on Mobile" era, mobile networks must meet requirements more diverse than ever. These requirements can be identified from three dimensions: number of connections, latency, and throughput. These three dimensions together will bring unprecedented challenges to future 5G networks:

1. Number of connections. Although a 4G network provides thousands of connections for each cell, a 4G network cannot meet the connection needs of Everything on Mobile. A 5G network provides up to a million connections per square kilometer. This will bring an exponential increase in the number of connections.

2. Latency. The latency on a 4G network, 50 ms, is half of that of a 3G network. However, applications such as self-driving cars still require much lower latency than a 4G network.

3. Throughput. A higher throughput will better meet consumer needs. The throughput of a 4G network is 10 times higher than that of a 3G network, but once 4K video services become popular, the 4G network cannot meet the new throughput demands. To meet the preceding requirements, 5G should have the following performance advantages over existing mobile communication technologies:

a) 100 billion connections
b) 1 ms latency
c) 10 Gbps throughput

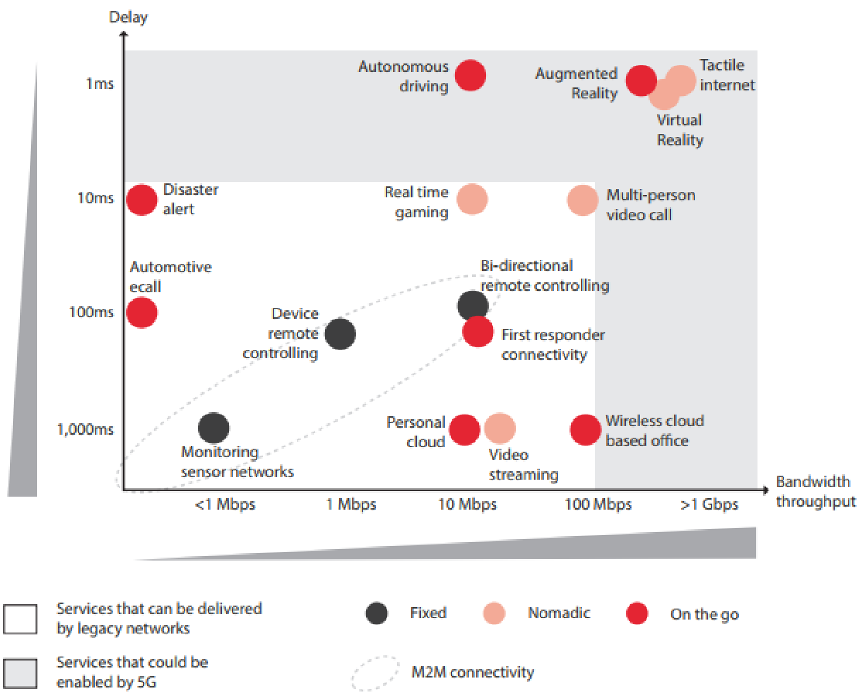


Figure 4-4 1Bandwidth and latency requirements of potential 5G use cases

Based on the features of 5G networks and the development trend for Internet of Things, Huawei has identified some typical application scenarios during the process of 5G innovation as below:

[1] 100 Billion Connections
[2] Personal Assets on Mobile

As wearable devices gain popularity, an increasing number of wearable technologies will be connected to the network. Wearable devices will provide healthcare management, improve quality of life, and work efficiency. For example, ultra-light, ultra-thin, low energy-consumption, and waterproof sensors can be implanted into sportswear. These sensors will monitor the atmospheric pressure, temperature, humidity, and air quality of the external environment, and monitor a wearer's blood pressure, heart rate, temperature, breathing, and skin humidity. The sensors will send the collected data to a database through a network, where a life management system analyzes the data and sends real-time messages to wearers to avoid putting their health at risk. The life management system also helps the wearer adjust diet, sleep, and exercise.

### 4.3.1 Smart Logistics

The capability of massive connection provided by 5G will enable extremely meticulas management for the logistics industry in the future. Before every piece in a delivery leaves the consighnor, a passive RFID will be attached to it. Throughout the entire the process of delivery, these RFIDs will report in real time the detailed information of the current position and environment of the delivered pieces they are attached to, such as temperature and humidity, light intensity, purity of air, speed of movement, intensity of vibration. With such data reported, the carrier and its clients can easily grasp the current positions and status of all deliveries, based on which, the carrier and its clients can plan the activities to take after the delivery scientifically. Also, once any alarm of exception is reported, the carrier can take countermeasures immediately and precisely.

### 4.3.2 Intelligent Agriculture

Agriculture applications in the future will realize the importance of managing and fine-tuning rural works. Farms of the future will widely deploy a large number of soil quality sensors providing real-time data on the amount of fertility and humidity in the soil to precisely control fertilization and irrigation. Near the farm, sensors will be deployed to monitor the temperature, humidity, wind, and sunshine. In the river, sensors will be deployed to monitor water quality. Timely corrective measures can be implemented if the data indicates that the river is polluted. Farmers can monitor diseases, blood pressure, and body temperature, as well as the locations of their livestock by implanting sensors into the animals. The intelligent agricultural system consisting of various sensors will reduce cost, improve efficiency, and increase needs for narrowband connections. If intelligent traffic management becomes a reality in future cities, more and more self-driving cars will appear on the road. To guarantee traffic safety, when a control command, braking for example, is sent to a car, the car must receive the command within 1 ms.

The latency of a 4G network cannot meet this requirement. With the latency of 4G network, a car driving at 100 km/h still moves 1.4 m from the time it finds an obstacle to the time when the braking command is executed. Under the same condition, with the latency on a 5G network, the car will move just 2.8 cm, and this performance is comparable with the standard of an anti-lock braking system (ABS).

### 4.3.3 10 Gbps Throughput

Virtual reality and immersive experience bring dramatic changes to many industries, including gaming, education, virtual design, healthcare, and art. Take education for example. For students living in areas with limited educational resources, virtual reality technology enables instruction and interaction with teachers in a virtual classroom, and even enables them to perform tasks such as carrying out experiments in a virtual laboratory.

To make this come true, the resolution of virtual reality image and immersive video needs to approximate to the amount of detail the human retina can perceive. This requires that the throughput be 300 Mbps and above, almost 100 times higher than the current throughput supporting HD video services.

Latency and throughput are two critical performance metrics of a communication network. Recently, a lot of attention has been focused on improving throughput (or spectral efficiency) of Wireless Wide Area Networks (WWANs) through the use of physical and MAC layer techniques, such as higher order modulation, MIMO and aggregation of bandwidth (multi-carrier). While some data applications directly benefit from the higher data rates, for many applications high data rates do not translate to improved user experience unless the latency is low. In this paper, we examine the Control plane (C-plane) and User plane (U-plane) latencies in an HSPA data system. We show that significant C-plane latency reduction can be achieved in HSPA by carrying signaling on HS-DSCH and E-DCH channels (as opposed to dedicated channels, which is the practice now). We also compare the C-plane and U-Plane latencies of HSPA and LTE, which have comparable spectral efficiency.

3rd Generation Partnership Project (3GPP) has standardized Code Division Multiple Access (CDMA) based packet-switched air-interfaces for downlink and uplink, called High-Speed Downlink Packet Access (HSDPA) and High-Speed Uplink Packet Access (HSUPA) respectively, together referred to as High Speed Packet Access (HSPA) [2][4]. In today's HSPA deployments, signaling messages are typically carried on dedicated channels. These dedicated channels are typically low rate channels (3.4 kbps), which allows larger spreading factors to be used so that more users can share the cell's resources. Since some signaling messages tend to be fairly large, the use of these low rate dedicated channels leads to call setup latencies that are in the range of a few seconds. Newer HSPA deployments are considering carrying signaling on HSPA channels (i.e., HS-DSCH on the downlink and E-DCH channels on the uplink). Such configurations not only allow better statistical use of power and code resources, but also allow signaling to benefit from higher rate HSPA channels. We will show that significant improvements to call setup latency can be achieved by carrying signaling on HSPA channels.

In parallel with the evolution of HSPA, 3GPP has also standardized an Orthogonal Frequency Division Multiplexing (OFDM) based air-interface, called Long Term Evolution (LTE) [3].

Though the spectral efficiencies of the two systems are comparable, Rel-9 LTE provides higher peak rates (due to its wider 20MHz bandwidth) compared to Rel-8/Rel-9 HSPA, which allows a User Equipment (UE) to receive on 10 MHz bandwidth with carrier aggregation. This paper focuses on the *call setup* and *user plane* latencies of the two systems. *Call setup latency* determines how quickly the user can start to receive service, while *user plane latency* is important since the performance of applications such as web browsing is very sensitive to the Round Trip Time (RTT) to the server.

Long-Term Evolution (LTE) is a new standard specified by 3GPP for fourth generation (4G) wireless communications. LTE provides high spectral efficiency, high peak data rates, short round trip time, and frequency flexibility. It relies on the following technologies; Orthogonal Frequency Division Multiplexing (OFDM) and Multiple-Input and Multiple- Output (MIMO), robust channel coding, scheduling and link adaptation [2]. LTE is interoperable with widely used technologies such as GPRS, WCDMA and HSPA, and this enables mobile operators deploying LTE to provide a seamless service and multimode devices for customers. Some companies have already launched commercial LTE networks, e.g., Verizon Wireless in the United States and Vodafone in Europe [4].The LTE technology has evolved over multiple releases which have led to improved data throughput, lower latencies, and increasingly flexible configurations . Its first release, R8, provided peak rates of 300 Mbit/s downlink and 75Mbit/s uplink, a radio-network delay of less than 5 ms, bandwidth sized in 1.4, 3, 5, 10, 15, or 20MHz blocks to allow for a variety of deployment scenarios a significant increase in spectrum efficiency compared to previous cellular systems (Universal Mobile Telecommunications System -UMTS and High-Speed Packet Access - HSPA).

Release 9 introduced some new service features and network architecture improvements such as evolved multimedia broadcast and multicast service, location services and dual layer beam forming. Release 10, LTE Advanced, features peak data rate of 3 Gbps in downlink and 1.5 Gbps in uplink, higher order MIMO antenna configurations supporting up to 8×8 downlinks and 4×4 uplinks, and Carrier Aggregation (CA). Release 11 and beyond includes enhancements to Carrier Aggregation, MIMO, and relay nodes, introduction of new frequency bands, and coordinated multipoint transmission and reception. LTE supports time and frequency division duplex Schemes (TDD/FDD) in the same frequency bands as those allocated to UMTS: 15 frequency bands (1 to 14 and 17) are FDD and 8 bands (33 to 40) are TDD [6]. LTE also supports three modulation schemes, which are QPSK, 16-QAM and 64-QAM. Error Vector Magnitude (EVM) is a parameter used to measure the quality of modulation. The minimum requirements for the EVM are 17:5% for QPSK, 12:5% for 16-QAM and 8% for 64-QAM. LTE downlink transmission scheme is based on Orthogonal Frequency Division Multiple Access (OFDMA), while the uplink transmission is based on Single Carrier Frequency Division Multiple Access (SC-FDMA).

The main drawback of OFDMA over SC–FDMA is its high Peak to Average Power Ratio (PAPR). OFDMA allocates individual users in the time and the frequency domain and its signal generation in the transmitter is based on the Inverse Fast Fourier Transform (IFFT).

OFDMA converts the wide-band frequency selective channel into a set of many fading sub-channels, which enables optimum receivers to be implemented with reasonable complexity during MIMO transmission [9]. In the uplink, SC–FDMA is more power efficient than OFDMA due to its low PAPR, which leads to decrease in linearity requirements and enhances the efficiency of the power amplifiers of the User Equipment (UE). SC-FDMA is based on discrete Fourier transform (DFT)-precoded OFDMA. Concerning uplink, LTE output power limits vary depending on the frequency band. The maximum TX power is 23 dBm.

### 4.3.4 Downlink and uplink

LTE downlink (from tower to device) transmission is based on OFDMA. The LTE downlink physical resources can be represented by a time-frequency resource grid. Resource elements are grouped into Resource Blocks (RBs) and each RB consists of 12 subcarriers with a spacing of 15 kHz in the frequency domain and 7 consecutive OFDM symbols in the time domain. The number of available RBs in the frequency domain varies depending on the channel bandwidth, and channel bandwidths may vary between 1.4 MHz and 20 MHz.

The transmitter and receiver structure of PDSCH in the physical layer starts with the grouped resource data which are in the form of transport blocks. PDSCH is used to transmit the Downlink Shared Channel (DL-SCH). The DL-SCH is the transport channel used for transmitting downlink data (a transport block). One or two coded transport blocks (code words) can be transmitted simultaneously on the PDSCH depending on the precoding scheme used. According to the processing steps of transmitting downlink data in PDSCH are given below:

- **Transport block CRC attachment:** A cyclic redundancy check (CRC) is used for error detection in transport blocks. The entire transport block is used to calculate the CRC parity bits and these parity bits are then appended to the end of transport block.
- **Code block segmentation and CRC attachment:** In LTE, a minimum and maximum code block size is specified so the block sizes are compatible with the block sizes supported by the turbo interleaver. Minimum code block size is 40 bits and maximum code block size is 6144 bits. The input block is segmented when the input block is greater than the maximum code block size.
- **Channel coding:** The channel coding scheme for PDSCH adopts Turbo coding, which is a robust channel coding. The coding rate of turbo encoder is 1/3.The code blocks undergo turbo coding which is a form of forward error correction that improves the channel capacity by adding redundant information. The turbo encoder scheme uses a Parallel Concatenated Convolutional Code (PCCC) with two recursive convolutional coders and a contention free Quadratic Permutation Polynomial (QPP) interleaver.

- **Rate Matching:** The main task of the rate matching block is to create an output bit stream to be transmitted with a desired code rate. As the number of bits available for transmission depends on the available resources the rate matching algorithm is capable of producing. The three bit streams from the turbo encoder are interleaved followed by bit collection to create a circular buffer. Bits are selected and pruned from the buffer to create an output bit stream with the desired code rate. The Hybrid Automatic Repeat Request (HARQ) error correction scheme is incorporated into the rate-matching algorithm of LTE.
- **Code Block Concatenation:** In this stage, the rate matched code blocks are concatenated back together. This task is done by sequentially concatenating the rate-matched blocks together to create the output of the channel coding.
- **Scrambling:** The code words are bit-wise multiplied with an orthogonal sequence and a UE-specific scramble LTE Radio Access Network (RAN) is comprised of the following protocol entities: Packet Data Convergence Protocol (PDCP), Radio Link Control (RLC), Medium Access Control (MAC) and The Physical layer (PHY).

The PHY transfers information to and from the MAC layer using transport blocks that convey data for at most two subframes.The PHY handles coding and decoding, modulation and demodulation, and antenna mapping. The LTE PHY is a highly efficient means of conveying both data and control information between an enhanced base station (eNodeB) and mobile User Equipment (UE) LTE System Toolbox functions is used to show the PUSCH throughput of transmit/receive chain. Channel noise is added to the received waveform which is then SCFDMA demodulated, resulting in a received resource grid for each receive antenna. Channel estimation is performed to determine the channel between each transmit/receive antenna pair. PUSCH data is then extracted and decoded from this recovered resource grid.Throughput for 10 frames (a) and for 20 frames (b). For 10 frames, the throughput is above 70% when SNR is -2.2dB and above. Also, the throughput is steady when SNR is -1.2dB and above. For 20 frames, the throughput is aboven70% when SNR is -2.2dB and above. Then the throughput Increases and becomes steady when SNR is 3.4dB.

Table 4. 1PDSCH Transmit diversity throughput simulationconfiguration:

| | |
|---|---|
| Code word | Single |
| Transmission Scheme | Transmit Diversity |
| Number of Transmitters | 4 |
| Number of Receivers | 2 |
| Multi-antenna Correlation | Medium |
| Propagation Channel | Extended pedestrian A (EPA 5 ) |
| HARQ | 8 HARQ retransmission process |
| Reference Measurement Channel( RMC) | R.12 |
| Number of Frames | 10 (first simulation) & 20(second simulation) |
| SNR range | [-5.8,-4.6,-3.4,-2.2,-1.2,0.2,1.2,2.2,3.4,4.6] |

## 4.4 Latency in TDD system

Most traffic can be assumed to be user or device-initiated, independently on whether the data is generated or consumed by the user equipment (UE). In a scheduled system, several TDD cycles are required for delivering even one scheduled round trip transmission of control or data signaling, such as request-response or data-acknowledgement pair. illustrates UE initiated data reception / transmission procedure in a scheduled system, requiring 4 TTD cycles from signaling perspective: one TDD cycle for the request signal in UL, possibly one TDD cycle for resource assignment signaling in DL and at least one TDD cycle for the actual data transmission, either UL or DL, followed by the corresponding acknowledgement. Consequently, the total air interface latency for this procedure is hard limited by the minimum enabled UL/DL switching time, often restricted by a certain UL/DL switching period. This leads to the requirement of flexible and fast link direction switching and increases the importance of short TDD switching guard times between the link directions.

Deployment densification with smaller cell sizes and utilization of higher carrier frequencies with larger bandwidth provide remarkable enablers for reducing air interface latency. Smaller cell sizes lead to decreased propagation losses, while delay spread may also be expected to decrease when moving towards utilization of higher carrier frequencies with larger bandwidth. The characteristics of LA environment together with certain expected improvements in component technology within the 5G timeframe, such as shorter TDD link direction hardware switching time, will eventually provide possibility for a TDD numerology optimized for dense deployment, e.g. shorter cyclic prefix (CP) and guard period (GP) times compared to the existing systems. Since the overhead from the guard times becomes significantly smaller, this new 5G optimized numerology further enables shorter frames and more frequent link direction switching.

Further, short OFDM symbols correspond to large subcarrier spacing, which is robust to the increased oscillator phase noise at higher carrier frequencies. These properties can further be utilized to design a 5G dense deployment optimized physical frame structure. In [4], we proposed a bi-directional control structure for each TDD frame, enabling opportunity to transmit/receive scheduling information (request/grant) together with synchronization signaling in every frame. The principle is shown in Figure 2. In [4], it was also demonstrated that very short frame lengths, such as 0.25 ms, are feasible with the proposed frame structure from the guard and control overhead point of view, using relatively low 60 kHz subcarrier spacing.

The demand for frequent link direction switching and the consequent need for short TDD guard times sets demands on the modulation methods suitable for a TDD-based air interface for 5G dense deployment. From latency perspective, it would be beneficial to adopt waveforms having good time localization, such that the aforementioned time overheads can be efficiently reduced. In that sense, orthogonal frequency division multiplexing (OFDM) modulation is well known to fulfill such time localization requirement due to the usage of the CP. This is an important benefit compared to filter-bank multi-carrier (FBMC) based waveforms, that are well-localized in frequency but correspondingly dispersed in time dimension. Consequently, transmission of a frame consisting of FBMC symbols is subjected to pre- and post-tails, leading to the need of longer TDD guard times between the link directions. These tails could be shortened but only with a cost of spectral re-growth [5], ultimately leading to the conclusion that from the latency perspective, it is justified to assume the OFDM waveform (and its enhancements) to be the most suitable basis for a 5G air interface designed for a dense deployment environment.

## *4.5 Modulation*

Diversity, the ability to exploit channel variations in time, frequency, and space for communication robustness, is the most powerful tool in the physical layer for achieving high reliability communications in a fading channel. For wireless communications, a Rayleigh fading channel represents the most challenging case in terms of achieving high reliability due to large fading dips. The cumulative distribution functions (CDF) of fading gains in Rayleigh fading channels of various diversity orders are shown in Fig. 1. It can be seen that without diversity, fading gains of -90 dB or lower occur with 10-9 probability. To mitigate fading, a 90 dB margin is thus needed for guaranteeing lower than 10-9 probability of fading-induced outage, which contributes to packet error events. With diversity orders 8 and 16, the needed margins reduce significantly to 18 dB and 9 dB, respectively.

Diversity may be achieved via spatial diversity using multiple transmit and receive antennas, via frequency diversity using multiple resource blocks of independent fading coefficients, and via time diversity using time slots of independent fading coefficients. For the targeted very low latency requirement, it is however difficult to exploit time diversity. Note that receive diversity not only achieves a diversity gain, but also provides a receiver processing gain due to coherent combining of the desired signal. The receiver processing gain results in a higher receiver signal-to-noise ratio (SNR) after coherent combining.

To further increase diversity, frequency diversity can be exploited by mapping the coded bits to multiple resource blocks that are sufficiently separated in frequency to have independent channel coefficients. According to [8], channel delay spreads in indoor industrial environments can be up to 300 ns, which implies that a coherence bandwidth of 3 MHz or more can be expected.

Required signal-to-noise ratio (SNR) for achieving packet error rate of 10-9 in channels of various spatial diversity orders are shown in Fig. 2, where 100 information bits are transmitted with a 100 μs transmission time interval with a rate-1/2, constraint-length 7 convolutional code and QPSK, which requires 1.5 MHz bandwidth, if 30% of the signal is assumed to be overhead. In the cases of having more than two transmit antennas, Alamut code is used across a pair of transmit antennas and FEC coding is used across antenna pairs to fully exploit diversity. This however results in a small SNR penalty in the rate-1/2 coding case compared to the case where each coded bit has full diversity. For example, the 4x1 case has a higher required SNR than the 2x2 case excluding the receiver processing gain. As illustrated, diversity reduces the required SNR significantly. A system with 8 base station antennas and 2 device antennas needs an SNR of 11.0 dB and 2.7 dB.

Though many modern communication systems (HSPA, LTE, 802.11/AC) use Turbo or LDPC codes as FEC for data [9-10], it is preferred to use convolutional codes in the low latency and high reliability MTC use case. Convolutional codes have similar performance as Turbo and LDPC codes for block lengths that are typical for this use case (e.g., up to a few hundred bits). In contrast to convolutional codes, Turbo and LDPC codes may for certain configurations have an error floor [11-12] that make these codes less efficient when the packet error rate shall reach very low levels (e.g. 10-9). Considering latency, convolutional code decoding has a shorter delay than the iterative decoder typically used for Turbo and LDPC decoding. This is partly due to lower decoding complexity.

But another important property of convolutional codes is that the decoder can process the code block while it is being received, and obtain decoded bits with a very small delay. This requires that interleaving is only performed over frequency, not over time. For control channels that have block lengths lower than 10 bits, block codes are preferred due to better performance and manageable decoding complexity.

Another important use of coding is to harvest diversity. As discussed above, diversity is a powerful tool for achieving high reliability, and to achieve transmit and frequency diversity in an OFDM system, it is essential to spread the coded bits over different diversity channels. Ideally, if the correct and erroneous code words differ in d positions, it is desired that these d positions are mapped to independent frequency bins or transmit antennas [13].

If a deployment has M diversity channels, the code rate needs to be low enough to have free distance (convolutional codes) or minimum Hamming distance (block codes) sufficiently larger than M. Code rate selection should also take modulation selection into account. Modulation and coding selection impacts both the required received signal power and the required bandwidth. In general, higher modulation order and code rate require additional signal power, but reduces the needed bandwidth, while low modulation order and code rate require more bandwidth but lower received signal power. When it comes to modulation, there are practical limitations such as transmitter and receiver impairments which typically limit the highest modulation order.

For example, in LTE the modulation with the highest modulation order is 256-QAM [14]. Thus, in our study, higher order modulations only include modulations up to 256-QAM, which together with a highest code rate of 1 sets a fundamental limit on the minimum bandwidth needed for low latency communication, independent of signal power. The selection of code rate and modulation should also take into account the above-mentioned requirements on code rate for transmit diversity and frequency diversity gains.

The required S/N0 for modulation schemes up to 256-QAM and code rates 1, 1/2, and 1/3, for a setup with one transmit antenna and eight receive antennas, where a 100 bit packet is transmitted in 100 μs over a frequency-flat Rayleigh fading channel, and where 30% of the signal is overhead. Here, S represents the received signal power and N0 is the one-sided noise power spectral density. Thus, S/N0 is a ratio between power and power spectral density, which can be related to the signal-to-noise power ratio by S/N0 = (S/N) B, where B is the bandwidth of the desired signal and N is the power of the noise, N = N0B.

# Chapter 5

# Low Latency 5G Architecture

## 5.1 Introduction

Typically in current LTE networks, it takes approximately 25~40 milliseconds (one-way) for a given packet to travel from UE to P-gateway before reaching the Internet. Taking a closer look at each network node and link inducing additional latency with break-down of network segments, we identify the following three key directions for reducing end-to-end latency in 5G network architecture: 1) new radio access networks (incl. latency reduction techniques in LTE), 2) distributed/flat network architecture, 3) intelligent end-to-end network orchestration with unified and converged transport networks.

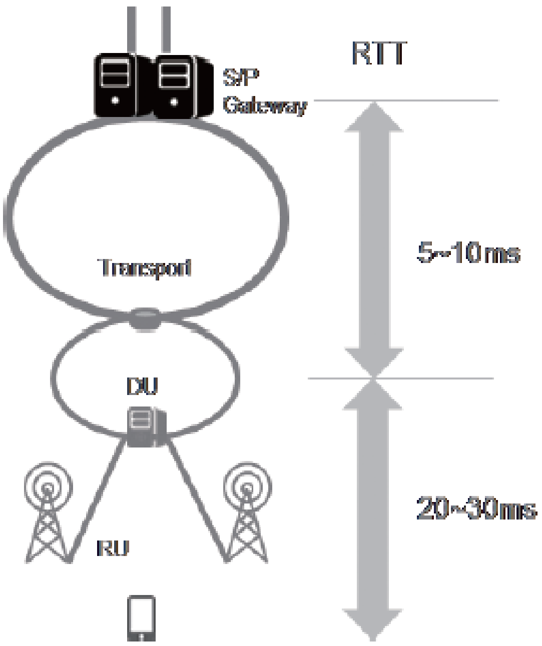


Figure 5-1 Network path and latency break-down in LTE-SAE architecture

Latency originated from radio access networks takes a dominant portion in end-to-end network latency. The latency requirement for 5G new radio access technology is still under discussion. However there seems to be a consensus on target latency of 1ms for 5G radio access network amongst operators and vendors. To achieve this, many research organizations and equipment manufacturers are currently looking into different aspects of the existing LTE technologies and new radio access technologies. Shorter TTI (transmission time interval) than 1ms of LTE system has been investigated while keeping the backward compatibility to the current LTE in 3GPP. Yet, LTE has a fundamental limitation that its TTI cannot be reduced below the OFDM symbol length. Therefore, 5G new RAT is expected to have shorter symbol length as well as shorter TTI than LTE to ensure lower latency in radio layer. In addition, flexible duplex and new frame architecture have been extensively investigated for additional latency reduction.5G RAT (Radio Access Technology) frame architecture that allows shorter physical layer round trip time with fast control signaling and fast TDD (Time-Domain Duplex) data switching periodicity.

Radio protocol optimization is another approach for reducing radio latency. In current LTE networks, when a given UE has data in buffer to send in uplink, it first sends scheduling request (SR) to eNB and needs to wait for grant from eNB, If this process can be enhanced or even removed (by, e.g. uplink pre-grant), radio layer latency may be further reduced. To discuss L2 protocol enhancement in LTE and make a LTE standard amended, a study item for latency reduction techniques was recently approved in 3GPP RAN2 and now underway (it also consider backward-compatible TTI reduction in LTE)[9].

Not all 5G radio access technologies improve latency. To boost up capacity in radio link, higher frequency such as millimeter-wave has been considered due to its high availability of broadband spectrum more than 100MHz. However, it suffers from higher atmospheric loss and higher constraint of NLOS (non-line-of-sight) while it allows more antenna to be integrated with the same form factor. Due to hardware complexities and power consumption issues, Analogue beam forming has typically been used for millimeter-wave, which cannot help sacrificing flexibility with less degree of freedom in intelligent beam searching as opposed to digital beam forming. This may induce increase in radio latency. For example, random access procedure needs more time to complete because PRACH needs to be transmitted multiple times to find best beams both in UE and in eNB In a higher mobility environment, beam searching procedure needs to be done more often, which would be another factor to increase radio latency in 5G. Therefore, careful design of beam forming architecture (e.g. hybrid beam forming architecture) and protocol (e.g. hierarchical beam searching) are indispensable for latency reduction in millimeter-wave 5G radio access networks.

## *5.2 Spectrum for 5G*

In order to support increased traffic capacity and to enable the transmission bandwidths needed to support very high data rates, 5G will extend the range of frequencies used for mobile communication. This includes new spectrum below 6GHz, as well as spectrum in higher frequency bands. Specific candidate spectrum for mobile communication in higher frequency bands is yet to be identified by the ITU-R or by individual regulatory bodies. The mobile industry remains agnostic about particular choices, and the entire frequency range up to approximately 100GHz is under consideration at this stage, although there is significant interest in large contiguous allocations that can provide dedicated and licensed spectrum for use by multiple competing network providers. The lower part of this frequency range, below 30GHz, is preferred from the point of view of propagation properties.

At the same time, very large amounts of spectrum and the possibility of wide transmission frequency bands of the order of 1GHz or more are more likely above 30GHz. Spectrum relevant for 5G wireless access therefore ranges from below 1GHz up to approximately 100GHz, as Figure 3 shows. It is important to understand that high frequencies, especially those above 10GHz, can only serve as a complement to lower frequency bands, and will mainly provide additional system capacity and very wide transmission bandwidths for extreme data rates in dense deployments. Spectrum allocations at lower bands will remain the backbone for mobile-communication networks in the 5G era, providing ubiquitous wide-area connectivity. The World Radio Conference (WRC)-15 discussions have resulted in an agreement to include an agenda item for IMT-2020, the designated ITU-R qualifier for 5G, in WRC-19. The conference also reached agreement on a set of bands that will be studied for 5G, with direct applicability to NX. Many of the proposed bands are in the millimeter wave region and include: > 24.25GHz to 27.5GHz, 37GHz to 40.5GHz, 42.5GHz to 43.5GHz, 45.5GHz to 47GHz, 47.2GHz to 50.2GHz, 50.4GHz to 52.6GHz, 66 GHz to 76GHz and 81GHz to 86GHz, which have allocations to the mobile service on a primary basis; and > 31.8GHz to 33.4GHz, 40.5GHz to 42.5GHz and 47GHz to 47.2GHz, which may require additional allocations to the mobile service on a primary basis. The mobile industry will strive to gain access to spectrum in the 6GHz to 20GHz range, but the policy directions being followed by regulators seem to be focused on frequency bands above 30GHz. In the US, the FCC has issued two Notices of Public Rule Making (NPRM) on bands above 24GHz. Ofcom has likewise indicated a preference for bands above 30GHz within the mobile industry. The capacity needs of the mobile industry will continue to be served by licensed spectrum, although novel sharing arrangements for spectrum will become progressively more important as restricted opportunities for new spectrum start to impact incumbent services such as satellite communication and radio location. Two examples of sharing arrangements include LSA planned in Europe for the 2.3GHz band and the Citizens Band Radio Service for 3.5GHz in the US.

## *5.3 5G technology components*

Beyond extending operation to higher frequencies, there are several other key technology components relevant for the evolution to 5G wireless access. These components include access/backhaul integration, device-to-device communication, flexible duplex, flexible spectrum usage, multi-antenna transmission, ultra-lean design, and user/control separation.

### 5.3.1 ACCESS/BACKHAUL INTEGRATION

Wireless technology is already frequently used as part of the backhaul solution. Such wireless-backhaul solutions typically operate under line-of-sight conditions using proprietary radio technology in higher frequency bands, including the millimeter wave (mmW) band. In the future, the access (base-station-to-device) link will also extend to higher frequencies. Furthermore, to support dense low-power deployments, wireless backhaul will have to extend to cover non-line-of-sight conditions, similar to access links. In the 5G era, the wireless-access link and wireless backhaul should not therefore be seen as two separate entities with separate technical solutions. Rather, backhaul and access should be seen as an integrated wireless-access solution able to use the same basic technology and operate using a common spectrum pool. This will lead to more efficient overall spectrum utilization as well as reduced operation and management effort.

### 5.3.2 DIRECT DEVICE-TO-DEVICE COMMUNICATION

The possibility of limited direct device-to-device (D2D) communication has recently been introduced as an extension to the LTE specifications. In the 5G era, support for D2D as part of the overall wireless-access solution should be considered from the start. This includes peer-to-peer user-data communication directly between devices, but also, for example, the use of mobile devices as relays to extend network coverage. D2D communication in the context of 5G should be an integral part of the overall wireless-access solution, rather than a stand-alone solution. Direct D2D communication can be used to offload traffic, extend capabilities and enhance the overall efficiency of the wireless-access network. Furthermore, in order to avoid uncontrolled interference to other links, direct D2D communication should be under network control. This is especially important for the case of D2D communication in licensed spectrum.

### 5.3.3 Flexible duplex

Frequency Division Duplex (FDD) has been the dominating duplex arrangement since the beginning of the mobile communication era. In the 5G era, FDD will remain the main duplex scheme for lower frequency bands. However, for higher frequency bands – especially above 10GHz – targeting very dense deployments, Time Division Duplex (TDD) will play a more important role. In very dense deployments with low-power nodes, the TDD-specific interference scenarios (direct base station-to-base-station and device-to-device interference) will be similar to the 'normal' base-station-to-device and device-to-base-station interference that also occurs for FDD. Furthermore, for the dynamic traffic variations expected in very dense deployments, the ability to dynamically assign transmission resources (time slots) to different transmission directions may allow more efficient utilization of the available spectrum. To reach its full potential, 5G should therefore allow for very flexible and dynamic assignment of TDD transmission resources. This is in contrast to current TDD-based mobile technologies, including TD-LTE, for which there are restrictions on the downlink/uplink configurations, and for which there typically exist assumptions about the same configuration for neighbor cells and also between neighbor operators.

### 5.3.4 FLEXIBLE SPECTRUM USAGE

Since its inception, mobile communication has relied on spectrum licensed on a per-operator basis within a geographical area. This will remain the foundation for mobile communication in the 5G era, allowing operators to provide high-quality connectivity in a controlled-interference environment. However, per-operator licensing of spectrum will be complemented with the possibility to share spectrum. Such sharing may be between a limited set of operators, or may occur in license-exempt scenarios. The Citizens Band Radio Service in the US in the 3.5GHz band and the 5GHz unlicensed spectrum are examples of managed and unlicensed sharing regimes respectively. New air interfaces like NX will likely be well served by more conventional licensed allocations of spectrum, mainly due to the need to establish a basic foundation for the technology to operate in an independent manner while interoperability is established with technologies like LTE. At some point, further allocations of spectrum for 5G may leverage the mobile industry's experience of sharing approaches in lower cellular bands.

### 5.3.5 MULTI-ANTENNA TRANSMISSION

Multi-antenna transmission already plays an important role in current generations of mobile communication and will be even more central in the 5G era, due to the physical limitations of small antennas. Path loss between a transmitter and receiver does not change as a function of frequency, as long as the effective aperture of the transmitting and receiving antennas does not change. The antenna aperture does reduce in proportion to the square of the frequency, and that reduction can be compensated by the use of higher antenna directivity. The 5G radio will employ hundreds of antenna elements to increase antenna aperture beyond what may be possible with current cellular technology. In addition, the transmitter and receiver will use beam forming to track one another and improve energy transfer over an instantaneously configured link. Beam forming will also improve the radio environment by limiting interference to small fractions of the entire space around a transmitter and likewise limiting the impact of interference on a receiver to infrequent stochastic events. The use of beam forming will also be an important technology for lower frequencies; for example, to extend coverage and to provide higher data rates in sparse deployments.

### 5.3.6 ULTRA-LEAN DESIGN

Ultra-lean radio-access design is important to achieve high efficiency in 5G networks. The basic principle of ultra-lean design can be expressed as: minimize any transmissions not directly related to the delivery of user data. Such transmissions include signals for synchronization, network acquisition and channel estimation, as well as the broadcast of different types of system and control information. Ultra-lean design is especially important for dense deployments with a large number of network nodes and highly variable traffic conditions. However, lean transmission is beneficial for all kinds of deployments, including macro deployments. By enabling network nodes to enter low-energy states rapidly when there is no user-data transmission, ultra-lean design is an important component in delivering high network energy performance. Ultra-lean design will also enable higher achievable data rates by reducing interference from non-user-data-related transmissions.

### 5.3.7 USER/CONTROL SEPARATION

Another important design principle for 5G is to decouple user data and system control functionality. The latter includes the provisioning of system information; that is, the information and procedures needed for a device to access the system. Such a decoupling will allow separate scaling of user-plane capacity and basic system control functionality. For example, user data may be delivered by a dense layer of access nodes, while system information is only provided via an overlaid macro layer on which a device also initially accesses the system. It should be possible to extend the separation of user data delivery and system control functionality over multiple frequency bands and RATs.

As an example, the system control functionality for a dense layer based on new high-frequency radio access could be provided by means of an overlaid LTE layer. User/control separation is also an important component for future radio-access deployments relying heavily on beam forming for user data delivery. Combining ultra-lean design with a logical separation of user-plane data delivery and basic system connectivity functionality will enable a much higher degree of device-centric network optimization of the active radio links in the network. Since only the ultra-lean signals related to the system control plane need to be static, it is possible to design a system where almost everything can be dynamically optimized in real time. An ultra-lean design combined with a system control plane logically separated from the user data delivery function also provides higher flexibility in terms of evolution of the RAT as, with such separation, the user plane can evolve while retaining system control functionality.

## *5.4 Device to everythingcommunication*

One way to effectively avoid this potential traffic bottleneck is to transform the existing centralized architecture into a more distributed architecture. For example, we can potentially place S/P gateways and have the traffic directly reach the Internet at the edge network closer to the user. , where future mobile network mainly consists of two types of clouds (i.e., Cloud RAN and Cloud Core). As opposed to today's LTE-SAE architecture where most of radio access functions are placed and run inside Cloud RAN, and most core functions (e.g., EPC) are placed and run inside Cloud Core, in 5G network architectures, the core functions may potentially be pushed out to the Cloud RAN. There are two main benefits of placing core functions at edge node where Cloud RAN is located. First, user packets can be directly sent out to the Internet, therefore reducing the overall latency. Second, as user packets are directly sent out to the Internet, the overall backhaul traffic being sent to the Cloud Core diminishes. This ultimately leads to the cost savings in mobile backhaul investments. Our preliminary experimental results showed an edge cloud would be able to reduce more than 30% of backhaul traffic with the help of mobile edge caching. There are various additional use cases and approaches to leverage this type of edge cloud, also referred as mobile edge computing discussed in ETSI MEC ISG and Fog computing. Edge cloud can also eliminate significant amount of signaling traffic by placing relevant control and even analytics functions close to devices and things. Other benefits of edge computing include higher level of security, as all signals and data traffic can be processed locally, and this is particularly attractive for public or B2B applications. Comparison between LTE-SAE architecture and flat/latency-optimized networks we have been designed for 5G network architecture. Detailed information on the 5G architecture that SK telecom has been developing can be found in there.

A lots of efforts have been done for the D2D (Device-to-Device) standardization in 3GPP Release 12 and we believe D2D will become increasingly widespread in 5G. This also implies 5G network becomes more distributed with D2D although cellular-assisted operation is somehow required from the viewpoint of QoS. For a given service. Analytics-engine will play an increasingly important role and can also be integrated with PCRF (Policy and Charging Rule Function) to promise intelligent operation of E2E networks.

The central orchestrator can also be used to optimize transport network path, via transport SDN (T-SDN) that can dynamically modify configurations of routers in transport networks. Looking at an individual optical router, different network layers from L1 to L2~L3 are now being integrated with advanced optical technologies, representatively called POTN (packet optical transport networks).this allows network switching as much of lower (optical) layer processing as possible for low latency services. IP layer switching usually takes 40~50 microseconds, on the other hand, optical layer switching takes less than 50 nanoseconds.

Technical challenges includes the limited flexibility for optical switching and wavelength allocations for beyond 100Gbps transmission and we expect elastic optical net-working technologies play important roles to get through these problems. Unified and converged dynamic transport networks realized by T-SDN and POTN can also reduce overprovisioning in backhaul/metro.

In order to provide service QoS, the network is typically overprovisioned and the resources used to implement the service are also sufficiently allocated such that QoS can be guaranteed even during the peak usage period. This has been the de-facto practice for guaranteeing QoS. However for 5G, this overprovisioning will no longer be a viable option as the peak usage is expected to be too high and ultimately lead to a cost prohibitive system. Therefore, the network and network functions must be designed to support policies for guaranteeing QoS even in the case of network failures and during the peak usage times. An end-to-end policy-based network service orchestrator as a key enabler that intelligently manages QoS and minimize latency dynamically in real-time even during the network is congested. More specifically, a central controller would consist of SDN controller (or Network OS) and NFV orchestrator interworking based on policies to dynamically change the network in order to provide QoS (e.g., minimum latency).

### 5.4.1 V2X Communication

As its name implies, vehicle-to-everything (V2X) communications and its solutions enable the exchange of information between vehicles and between vehicle network infrastructures. The goal of V2X is to improve road safety, increase the efficient flow of traffic, reduce environmental impacts and provide additional traveler information services. V2X communications consists of four types of communications: vehicle-to vehicle (V2V), vehicle-to-infrastructure (V2I), vehicle-to-network, (V2N) and vehicle-to-pedestrian (V2P). Based on extensive analysis of crash data from 2004 to 2008, the U.S. Department of Transportation (USDOT) has concluded that a fully implemented V2X system can address 4.5 million crashes, which is 81 percent of all multi-vehicle, unimpaired crash types. The USDOT is expected to mandate that vehicles manufactured in late 2019 and beyond deploy dedicated short range communication (DSRC) devices to support V2V and V2I communications.

The U.S. DSRC solution has gone through extensive testing by the auto ecosystem players for over 10 years and utilizes dedicated spectrum at 5.9 GHz. However, DSRC has several challenges. The system relies on road side units (RSUs), which are not currently deployed. Meanwhile, at the physical layer, several inefficiencies arise due to the asynchronous nature of the system, resulting in reduced performance, such as range. In the long run, there is no evolutionary path (or IEEE 802.11 standards activities) to enable improvements in the DSRC physical/MAC layers with respect to range, robustness and reliability. Fortunately, DSRC is not the only solution for V2X communications. LTE and fifth-generation (5G) cellular systems have the potential of supporting not only existing DSRC use cases, but also the more challenging and futuristic use cases that require low-latency, high reliability or high bandwidth. Cellular V2X could also could complement DSRC communications to enhance V2X communications capabilities. Mobile operators have the potential of providing additional value to the overall V2X solution. Cellular networks already blanket nearly the entire U.S., including rural interstates. Those networks could be used to distribute certificate and certificate revocation lists and for RSU backhaul communications.

Additionally, LTE networks can extend the V2X range from the 300 meters that DSRC can achieve to several kilometers or more. This extended range could benefit drivers by providing earlier notifications of accidents, road conditions and traffic congestion ahead. Automotive manufacturers have been investigating crash avoidance technologies for over 20 years. In September 2014, GM announced that Cadillac would launch a CTS sedan with V2V communication technology in 2017. Several auto manufacturers around the world are conducting V2X trials jointly with mobile operators and technology providers. No automotive manufacturers have made any public announcements on their plans to deploy V2X vehicles for the U.S. market.

In light of LTE evolution’s toward 5G, it is becoming evident that vehicles connected to the cellular network will be able to support superior V2X capabilities and possibilities. Viewed as a long-range new sensor, cellular V2X communications can enable new levels of automated driving. Cellular’s extensive coverage in rural, urban and suburban areas means V2X services have an enormous addressable market. That directly benefits every driver and passenger by enabling safer traffic flow and a more enjoyable travel experience. In summary, V2X communications are a critical component of the connected car of the future. Thus, the cellular ecosystem stakeholders should engage in early efforts to assess the forthcoming capacity and coverage needs of connected vehicles. A partnership between the cellular and automotive industries is critical for enabling a best-in-class vehicle connectivity solution.

V2X stands for Vehicle to Vehicle and Vehicle to Infrastructure communication. In Europe it’s also called Car to Car (C2C) or Car to Infrastructure (C2I) Another common name is DSRC (Dedicated Short Range Communication) Enables vehicles to communicate 2-way to other surrounding vehicles and to roadside units (each vehicle is broadcasting) Communication uses 5.9GHz, 802.11p based network V2X needs high quality position and time information.

Figure 5-2V2V and V2X overview

An overview of the main issues challenge for 5G systems, as driven by the fast changing mobile network evolution and the forecast expansion of use cases and applications.Since its inception, the use and role of cellular communications has been expanding with each generation of cellular technology. Cellular communications have evolved from simple voice communications to a broad array of voice, video and data. Cellular communications have also evolved from person-to-person communications to machine-to-machine (M2M) communications, including the telematics systems now used in many cars and trucks for applications such as remote diagnostics and automatic crash notification. The success of these telematics applications, along with 4G's extensive geographic coverage and the advanced capabilities of 5G, make cellular the ideal foundation for V2X.

The primary objective of V2X is to improve vehicular safety by reducing the number of vehicular crashes. The USDOT analyzed 2004-2008 crash statistics and concluded that a fully implemented V2X system can address 4.5 million crashes or approximately 81 percent of all multi-vehicle unimpaired crash types. This white paper describes the role and benefits that cellular V2X can provide to support the USDOT objectives of improving safety and reducing vehicular crashes. Cellular V2X can also be instrumental in transforming the transportation experience by enhancing traveler and traffic information for societal goals such as increased mobility and reduced pollution. Moreover, Cellular V2X can enable many convenience applications related to transportation. Cellular V2X is an umbrella term for 3GPP-defined V2X technologies, encompassing both LTE- and forthcoming 5G-based V2X systems. This white paper discusses the potential role of 3GPP solutions to support V2X and the added value that could be provided by cellular operators.

### 5.4.2 V2x USES

1. Safety, automated driving and advanced driver assistance systems (ADAS), which require high reliability, low-latency message transfer at high speeds. Examples are forward collision warning, emergency electronic brake light (EEBL), control loss

warning, blind spot and lane change warning, as well as vulnerable road user (VRU) safety applications.

2. Situational awareness, which entails high reliability and longer latency requirements, while still supporting high speed. Examples are queue warning and hazardous road condition warning.
3. Mobility services, encompassing communication to support intermodal travel, congestion reduction entailing support for devices that have intermittent connectivity, power constraints, and for more complex security if confidentiality is required. Examples are automated parking and tolling systems, traffic advisories and dynamic ride sharing.
4. Auxiliary services/comfort use cases, an umbrella term for use cases involving vehicles and having commercial value, require high data rates and flexible types of communications. Examples include infotainment, local information, route planning, map dissemination and fleet management. These services are also known as "personal mobility services".

### 5.4.3 The technologies used by V2X

The technologies used by V2X include traditional WAN (Wireless Access Network) and Wi-Fi communications as well as Wireless Access in Vehicular Environments (WAVE), which is based on DSRC for the lower OSI layers, and finally the emerging LTE-based V2X communications. In Europe, an equivalent to DSRC is the ETSI ITS-G5 standard. 1 For V2V in particular, while the lower layers may differ (e.g., in the radio signal waveform used or the spectrum channelization), the transport and especially the application layer enjoy some synergy. There are efforts to harmonize the transport protocols used in the U.S. with those developed in Europe, for example. The application layer messages are developed by car manufacturers in various standards bodies (e.g., SAE, ETSI-ITS) with input from consortiums (e.g., Car-to-Car Communication Consortium) and are very similar in structure, at least for the safety messages. DSRC, which is based on IEEE 802.11p, is an incumbent technology designed nearly two decades ago. It has undergone extensive standardization, product development and field trials by many stakeholders. There are strong proponents of DSRC in the current V2X ecosystem, particularly in the U.S. DSRC technology has been tested in numerous field trials, with the Ann Arbor Safety Pilot Model Deployment being the most widely known. Semiconductor companies such as Qualcomm, NXP Semiconductors, Renesas and Autotalks also designed and tested DSRC compliant products, and a U.S. automaker included a DSRC modem in its newest models. In Europe, there have been several ITS "plugtest" events organized by ETSI since 2011, as well as extensive field trials.

### 5.4.4 LTE-BASED V2X SERVICES

First-generation LTE V2X can and should deliver the same applications—and the same look and feel—of other V2X solutions, with the most significant difference being that LTE V2X offers better-performing radio access technologies. With better performance, LTE V2X solutions can broaden the breadth and effectiveness of currently envisioned V2X safety, mobility and environmental applications. To achieve this consistency in look and feel, LTE V2X solutions should re-use the message sets and many of the on-board performance standards envisioned for V2X. Moreover, the IEEE 1609.2 security services and framework, and the Public (PKI) system that enables trusted, anonymous safety message exchange with new vehicles, should be adapted to work with LTE V2X. This adaptation and adoption would capitalize on the accomplishments and consensus of the Intelligent Transportation Systems standards community and deliver the transformational low-latency applications that have been envisioned for nearly 20 years. This re-purposing of DSRC application and security services standards for LTE V2X make sense. Over the years, and with significant transportation stakeholder input, the SAE DSRC Technical Committee has developed a data dictionary, SAE J27355, that defines 16 (and counting) specific V2X messages.

For example, this standard includes the well-known basic safety message (BSM), which broadcasts what is essentially the vehicle state vector in order to provide by-lane target classification to enhance advanced driver assistance systems (ADAS). Another SAE J2735 message is the traffic signal phase and timing (SPaT), which enables broadcast of green-to-yellow-to-red transitions. Lesser known V2X messages include the traveler information message and emerging basic information or basic infrastructure messages. In fact, SAE J2735 contains over 230 data elements, enabling a wide range of customization options. Therefore, in the foreseeable future, combinations of the elements defined SAE J2735 can handle most envisioned use cases. In the more distant future, SAE J2735 will conceivably be enriched to accommodate the anticipated increased performance with LTE V2X (e.g., longer range, lower latency/higher frequency of transmission). Ultimately, it is reasonable to believe that the SAE standards community will in turn adapt and adopt LTE. It is reasonable to expect this cycle of adaptation and adoption to expand the scope of envisioned connected vehicle services, given that LTE V2X is not just short range, ad hoc message broadcast and reception. Because LTE V2X has V2V, V2I, VSP and V2N concepts, the types of messages and services enabled will likely transcend even those combinations available in SAE J2735. In fact, the very existence of V2N in combination with V2I and V2V enables additional participants, concepts and spectrum, to include potential use of existing cellular systems (V2N) in tandem with LTE V2X side link messages (V2V, V2I) or short range uplink and downlink (V2I). In the end, governments, private toll companies and other road owners/operators will be given tremendous flexibility because they can deliver for-the-public-good V2X safety messages, given that common spectrum is available. The road operator can additionally leverage spectrum owned by mobile operators to deliver V2N data exchange in providing traffic management services based on local or aggregated V2V and V2I messages. As a result, all actors win: The road operator’s transportation management system is enabled with a cost-effective LTE V2X implementation alternative, which the mobile operator delivers. Mobile operators can charge to serve the road operator with network equipment and can also charge individual users who realize value in using cellular spectrum for commercial traveler information or other V2N services. Ultimately consumers—drivers in this case—realize the value, as they receive important, high-quality pre-trip, enroute and safety-critical information via the same service.

### 5.4.5 V2X security

This section begins by focusing on the existing DSRC security as it is the most mature. The discussion then shifts to the emerging cellular LTE-based V2X system. Note that the security mechanisms specified for V2X by the Standards Development Organization (SDO) ETSI-Intelligence Transportation System (ITS) are largely like those specified by IEEE. This section ends by outlining a possible (likely) outcome of the security for the emerging cellular V2X specified by 3GPP.

4.5.1 REGARDING DSRC SECURITY DSRC security encompasses two aspects: the ability for a vehicle to verify that received safety messages are authentic, and; protecting the privacy of the vehicle/user. These basic requirements have also been recognized in other standards bodies (ETSI-ITS, 3GPP).

In the U.S., the privacy requirements are defined in the USDOT National Highway Traffic Safety Administration, DOT HS 812 014: Vehicle-to-Vehicle Communications: Readiness of V2V Technology for Application. 7 The ability for a vehicle to reliably receive correct Basic Safety Messages (BSMs) is crucial for the functioning of the safety application. BSMs contain the sender's accurate position speed, and other parameters. The security risk of false message injection by an attacker can be mitigated by requiring all senders to digitally sign each message. The cost of that security measure is an increase of the message size sent over the air, and increased computational load at the receiver side due to the high rate of signature verifications. A Public-Key Infrastructure (PKI) that distributes and manages digital certificates for vehicles is necessary for the security of DSRC communications. To this end, the IEEE 1609.2 standard defines a Security Credential Management System (SCMS), which implements a PKI with some additional features.

### 5.4.6 LTE V2X Security

LTE has the most advanced security solutions built into the standard itself by default, thanks to the extensive work of 3GPP over the years. 5G will continue in this paradigm by keeping with the same high-standards for the V2X domain. Today's 3GPP standards work includes developing capabilities that will benefit V2X, starting with enhancements in LTE to support V2X use cases. 5G will support more advanced use cases. Two requirements are particularly noteworthy:

1) The need for direct, ad-hoc, broadcast, secure communication without any a priori configuration of security by the network 2)
2) Management of identities for user privacy from the network or other third parties there are two types of LTE transport-level security mechanisms: LTE security protecting the UE signaling and communications with the LTE network, and LTE device-to device (D2D), a.k.a. ProSe, communications security. 8 LTE security uses a pre-shared symmetric keying for one-to-one data protection between UE and network. For user plane data, only confidentiality (encryption) is applied; there's no integrity protection on application layer messages exchanged with the network.

There is no integrity protection on the user data and, because the key is shared by all, no way to identify for sure which of the group members actually sent the data. Most importantly, pre-shared keys alone cannot provide some essential information about the V2V message's sender.

The receiver must have a way to verify that the received message was generated by a trusted entity. This is because for the V2V safety application, it is crucial to minimize false alarms. Confidentiality for safety messages is not needed, while strong integrity is paramount. Hence, neither LTE security nor Prosecurity are applicable to V2V. If V2X communications between one vehicle and the network can use the traditional UE to network LTE link, then the LTE security is applicable, as well as all signaling flows defined in current standards.

### 5.4.7 Currently used V2X services

In order to support increased traffic capacity and to enable the transmission bandwidths needed to support very high data rates, 5G will extend the range of frequencies used for mobile communication. This includes new spectrum below 6GHz, as well as spectrum in higher frequency bands. Specific candidate spectrum for mobile communication in higher frequency bands is yet to be identified by the ITU-R or by individual regulatory bodies. The lower part of this frequency range, below 30GHz, is preferred from the point of view of propagation properties. At the same time, very large amounts of spectrum and the possibility of wide transmission frequency bands of the order of 1GHz or more are more likely above 30GHz. Spectrum relevant for 5G wireless access therefore ranges from below 1GHz up to approximately 100GHz. Spectrum allocations at lower bands will remain the backbone for mobile-communication networks in the 5G era, providing ubiquitous wide-area connectivity. The World Radio Conference (WRC)-15 discussions have resulted in an agreement to include an agenda item for IMT-2020, the designated ITU-R qualifier for 5G, in WRC-19. The conference also reached agreement on a set of bands that will be studied for 5G, with direct applicability to NX. Many of the proposed bands are in the millimeter wave region and include: > 24.25GHz to 27.5GHz, 37GHz to 40.5GHz, 42.5GHz to 43.5GHz, 45.5GHz to 47GHz, 47.2GHz to 50.2GHz, 50.4GHz to 52.6GHz, 66 GHz to 76GHz and 81GHz to 86GHz, which have allocations to the mobile service on a primary basis; and > 31.8GHz to

33.4GHz, 40.5GHz to 42.5GHz and 47GHz to 47.2GHz, which may require additional allocations to the mobile service on a primary basis. The mobile industry will strive to gain access to spectrum in the 6GHz to 20GHz range, but the policy directions being followed by regulators seem to be focused on frequency bands above 30GHz. In the US, the FCC has issued two Notices of Public Rule Making (NPRM) on bands above 24GHz. Ofcom has likewise indicated a preference for bands above 30GHz within the mobile industry. The capacity needs of the mobile industry will continue to be served by licensed spectrum, although novel sharing arrangements for spectrum will become progressively more important as restricted opportunities for new spectrum start to impact incumbent services such as satellite communication and radio location. Two examples of sharing arrangements include LSA planned in Europe for the 2.3GHz band and the Citizens Band Radio Service for 3.5GHz in the US.

### 5.4.8 Broad variation of requirements and service characteristics

The main challenges for 5G system are the continued evolution of mobile broadband and the addition of new services e.g., massive sensor communication and vehicular to anything communication, requiring shorter setup times and delay, as well as reduced signaling overhead and energy consumption. Mobile broadband of the future will have significantly increased traffic volumes and data transmissions rates, but also many more use cases. They include not only traffic between humans and between human and the cloud, but also between humans,

Sensors, and actuators in their environment, as well as between sensors and actuators themselves. Some new key applications with disruptive characteristics follow project

Firstly, massive machine communications (MMC) is envisioned, whose main challenges are

i) To support 10100 times more devices than today;
ii) To allow very long battery lifetimes (on the order of 5+ years) of the wireless device;
iii) To incur minimum signaling overhead;
iv) To enable low cost wireless devices;
v) To support efficient transmission of small payloads with fast setup and low latency.

A final note is made on video streaming, which is already the biggest contributor to worldwide

Traffic today, at least in the fixed part of the Internet, and is expected to shift to mobile broadband connection as soon as the current technologies and billing plans will allow this. Moreover, the future video encoding and playback advances, including 3D, very high quality encoding, 4K resolution, and multiage, will further increase the capacity requirements.

Some examples of very diverse requirements for some use cases of business and social below:

Table 5. 1 diverse requirements for some use cases of business and social uses

| Requirement | Very strict | Intermediate | Relaxed |
|---|---|---|---|
| **High bit rate** | Video equipments | Typical applications on smart phones and tables | IoT, V2V |
| **Fast mobility** | Applications running on smart ohones and tablest on the raod | Everything else | Home and Office applications, IoT(most) |
| **High reliability** | PPDR,IoT (some),V2V/I | Everything Else | _ |
| **Low Latency** | Game consoles, IoT, V2V/I ,PPDR (some) | Web & mobile apps,cloud computing | IoT(some) |
| **Low energy consumptions** | IoT devices (most) | Smart ohones and tablets | Cabled devices |

## *5.5 Energy efficiency*

Classical designs for wireless communications, which tend to maximize rate, capacity and Coverage, potentially lead to solutions where energy efficiency drops. Energy efficiency is Understand from two points:

On the one hand, the energy spent by the infrastructure may increase, implying high operational costs for the operator that will indirectly affect also the invoice of the finals users.

On the other hand, some communication strategies require high computational burden at the terminal side having negative impact on battery lifetime.

Hence, theintelligent use of energy becomes a major new target in addition to the classical design criteria. Currently two approaches to reduce energy consumption on the radio link exist. First, small cells reduce the distance to the terminal. The main challenges of this approach are related to providing an economic backhaul solution and to minimize the additional deployment cost. The second approach is massive MIMO, where energy is more focused towards the user by means of more directive beams. In this way, less energy is wasted yielding interference for other users at the end. The challenges of massive MIMO include the diffusion of energy due to scattering in NLOS scenarios, limiting the achievable directivity, and the complexity of spatial multiplexing of users. Both in the terminal and at the base station, the goal of minimizing the energy consumption per bit will require a paradigm shift in wireless system design to dramatically improve efficiency in terms of power and spectrum usage. Further research on implementation technologies is necessary, focused on low power hardware architectures and energy efficient signal processing some approaches have been proposed on multihopCooperative networking, and wireless network coding [Car12].

There are further potential savings by operating the network with energy efficiency in mind. Nowadays base stations consume a constant power, regardless of the traffic load. During off-peaktraffic hours, small cells are switched off while coverage is maintained by macrocells. For active base stations serving a single user, following Shannon's theorem, the most energy efficientsituation would be to use the full bandwidth and to reduce power so that the throughput target is met. However, an interference limited multiuser scenario is more typical in mobile networks. Serving multiple users having different signaltointerference. Ratios in a TDMA fashion such as round robin, changing the power dynamically would result in unpredictable interference in adjacent cells. The same holds for OFDMA, implying inhomogeneous interference on different frequency sub bands.

Hence, current PHY and MAC layers design needs technology advances, including dynamic power control that is optimally coordinated among the users and with surrounding cells so that there is proportionality between the traffic and the energy consumption .There is a need for network architecture advances required to:

i) Include small cells and larger antenna arrays efficiently into the network design.

ii) Switch on/off base stations depending on the traffic load

iii) Achieve traffic proportionality at PHY layer.

Mobile devices with advanced capabilities such as smartphones or tablets may present Important requirements in terms of energy, not only as far as transmission is concerned (which depends for example on the data flows, the type of application or the wireless network topology) 4/18 but also regarding other components such as CPU, screen or audio devices at the user equipment. The offloading of applications today hosted by the mobile terminals towards the serving base stations or a (micro) data center may also contribute to energy efficiency. This way, the execution of resource hungry applications is shifted to processing elements that have more efficient computational and caching capabilities.

There is also a need to reduce energy consumption in the backhaul network, both in RAN and Core, in order to reduce network operational costs. Energy efficiency in the backhaul becomes increasingly critical as the access segment of the network consumes up to 90% of the totalTelecom network energy cost. Historically, this huge number is related to the use of copper; Withthe increasing use of optical fibre, the energy requirement is reduced. The access network has adistributed (tree) topology to aggregate the traffic. The enormous heterogeneity of fixed and wireless final drop technologies (i.e. e.g. FTTH, PON, AON, Wi-Fi, WiMAX, UWB etc.) makes economies of scale rather problematic. More unified and standardized fixed access solutions would allow much higher volumes, and thereby higher integration densities, much lower cost and reduced energy consumption. For instance, the use of an active remote node, originally put forward in Ethernet PONs [Chan10], was recently proposed as a common platform for fixed wireless convergence.

This node locates the network intelligence closer to the end usersand performs statistical multiplexing of traffic from fewer users, which allows to handle locally some traffic flows (such as the signaling between cooperative base stations), therefore reducing the backhaul load and enabling a more energy efficientoperation. Moreover, such lower levelaggregation requires less power-hungrycircuitry which, in turn, also makes it possible to use renewable energy sources only.

### 5.5.1 Network infrastructure

Small access nodes, with low transmit power and no precise planning requirements, are Conceived to be densely deployed, resulting in an Ultra Dense Network (UDN). This approach will improve spectral efficiency by reducing the distance between transmitters and receivers, and to improve macro cell service by offloading wireless traffic, thus freeing radio resources in the access. Network densification is a way to increase the capacity and datarate towards 2020. UDNs are a step further towards low cost, plugand play, self-configuring and self-optimizingnetworks. 5G will need to deal with many more base stations, deployed dynamically and in a heterogeneous manner, combining different radio technologies that need to be flexibly integrated.

Moreover, a massive deployment of small access nodes induces several challenges such as an adverse interference scenario or additional backhaul and mobility management requirements, which 5G needs to address. 3GPP is currently working on small cells solutions to reduce the intersite distance [3gpp36.932] but, at the time being, pilot contamination and interference still limit the possible densification. Different levels of coordination/cooperation among small cells are key to enhance the network capacity and keep interference at an adequate level, to manage mobility and spectrum, to ensure service availability and response to non-uniform Traffic distribution between neighboring access points.

With the increasing density of networks, also the backhaul will become more heterogeneous and possibly also scenario dependent (i.e., fibre, wireless backhaul or other non-idealtypes of backhaul might be used depending on their availability). In addition, the connectivity among the network nodes may change in order to allow for fast direct exchange of data between them (which will be challenging in ultra-densedeployments). The heterogeneous backhaul structure will also influence the operation of the radio access networks, e.g. latency differences on backhaul links will impact intercell coordination and cooperation algorithms.

Therefore, both radio access network and backhaul network need to be aware of limitations and capabilities of each other. This may for instance imply an extended SON applied to radio access networks which also uses information provided about the backhaul network.

The required flexibility of the network itself will require new concepts on network management inthe backhaul such as the application of Software Defined Networking (SDN) principles in order toachieve fast rerouting and congestion control, mainly in the access part . SDN concepts enable us to adapt the operation of the backhaul network to the needs of the radio access network.

For example, the selection of IP breakout anchor points may depend on the current backhaul traffic situation and QoS requirements in the radio access networks. Furthermore, the smaller the cells in the radio access network, the higher the temporal and spatial traffic fluctuations. This implies that also the backhaul network may experience a higher variance of traffic. Besides, current trends suggest that Infrastructure as a Service (IaaS) can be supported by small cells in order to offer innovative proximity services and to enable a series of advantages for end customers. With this approach, energyscarce. Capacity limited mobile devices can offload highly demanding computational tasks into proximal fixed units or use them for storage.

This entails that novel mechanisms are needed to efficiently allocate resources, understood in a wide sense (radio/computation/storage/energy), including contextual information metrics and clustering techniques for small cells. Another important aspect in the network infrastructure is related to the exposure of end users to electromagnetic field (EMF).There is today a public concern concerning EMF induced by wireless networks. By reducing the distance between receivers and transmitters, small cells enable the minimization of the power emitted by the mobiles phones and the total EMF exposure because, currently, the most important contribution is linked to the user equipment. 5G architecture combining small cells, heterogeneous networks and offloading should inherently enable minimizing the human EMF exposure.

Table 5. 2results in the delay calculations

| Delay Components | Delay Value |
|---|---|
| Transmission time uplink+Downlink | 2 ms |
| Buffering time ( 0.5 * transmission time | 2*.5=1 ms |
| Retransmission | 2*.1*8=1.6 ms |
| Uplink scheduling request | .5*5=2.5 ms |
| Uplink scheduling grant | 4 ms |
| UE delay estimated | 4 ms |
| EnodeB dealy estimated | 4 ms |

| | |
|---|---|
| Core Network | 1 ms |
| Total delay with pre allocated resource | 13.6 ms |
| Total delay with scheduling | 20.1 ms |
| | |

### 5.5.2 Ultimate goal

Table 5. 3 The ultimate goal

| **Delay Components** | **Delay Value** |
|---|---|
| Transmission Time Uplink+ Downlink | 2 ms |
| Buffering time (0.5 x transmission time) | 2*.5*1=1 ms |
| Retransmission (10%) | 2*1.8*1=1.6 ms |
| Uplink scheduling request | .465*5=2.31 ms |
| Uplink scheduling grant | 1 ms |
| UE delay estimated | 4 ms |
| eNodeB delay estimated | 4ms |
| Core network | 1 ms |
| Total delay with pre allocated resource | 13.6 ms |
| Total delay with scheduling | 15.91 ms |

# Chapter 6

# Proposal

## *6.1 Introduction*

The main physical channel present in LTE were physical downlink shared channel (PDSC) which carried data for all users. The physical downlink control channel (PDCCH) which contains all information required for UE to receive and transmit data to eNodeB. Latency measurements are done in all phases of the lifetime of a radio access network system; starting when verifying a new software release or system component, and continuing when deploying a system and after the system is put in commercial operation.

Since the introduction of LTE in 2009, several improvements had been developed, however, mainly targeting the increase of the maximum data rates of the system, e.g. Carrier Aggregation, 8×8 MIMO, etc. To also get full benefit of these data rate enhancements; we strongly believe that continuous enhancements of the latency of LTE should also be an important part of the future evolution track of LTE.

Although there has been some improvement on Low latency scheduling for 6TiSCH Networks. The 6TiSCH working group is standardizing the low-power wireless protocol stack for the Industrial IoT. The default scheduling function (SF0) standardized by 6TiSCH uses simple random slot selection. This paper proposes the Low Latency Scheduling Function (LLSF), a new scheduling function which daisy-chains timeslots rather than picking them randomly.

We implement LLSF in OpenWSN and evaluate its performance experimentally. LLSF yields 82.8% lower end-to-end latency on a 5-hop path than SF0, at no extra costs. The 6TiSCH working group is standardizing the low-power wireless protocol stack for the Industrial IoT. The default scheduling function (SF0) standardized by 6TiSCH uses simple random slot selection. This paper proposes the Low Latency Scheduling Function (LLSF), a new scheduling function which daisy-chains timeslots rather than picking them randomly. We implement LLSF in OpenWSN and evaluate its performance experimentally. LLSF yields 82.8% lower end-to-end latency on a 5-hop path than SF0, at no extra costs.

Moreover, there is a dedicated work done with the specific vehicular applications in mind, in order to enable the usage of the already existing infrastructure for the new use cases V2V, V2I, V2P. Rel'14 TR 22.885 gathers the use cases that the standard intends to support and defines the target latency for critical messages as 100 ms and lower [6]. Rel’14 TR 36.885 analyzes the latency requirements of the current 4G technology in order to introduce V2X communication via LTE: the current network performance is analyzed including Uu interface and the side link usage, and various improvement items are identified. Such effort is done in order to overcome the fluctuations of the LTE latency due to the network conditions or dimension, and to enable a guaranteed service.

So, there has been number of studies in order to achieve low latency in various ways. But still hypothesis for two control region has not been proposed yet to gain low latency that we will be discussing throughout our book.

## Proposed mechanism for low latency

### *6.2 Proposals:*

1. Dual Control Plane (DCP)
2. Prediction of uplink transmission

## Dual Control Plane:

The two control planes PDCCH and PUCCH may allocate resources to be used by the UE and eNodeB. The TTI is thus predefined by symbols predicted by the control channel. Present control region may indicate which PDCCH format should be used. Present control region will indicate which symbol in Control region 2 should use which RNTI. Control region 1 will allocate resources for some delay tolerant apps and leave some delay sensitive apps to be allocated by control region 2. Delay tolerant apps include text messaging, short voice calls whereas delay sensitive apps include V2V, V2I, and M2M etc. If Control region finds no data for delay sensitivity apps it can allocate for delay tolerant apps to avoid wastage.

## Prediction of uplink transmission:

Packets will be made assuming the patterns of previous uplink transmission packets. Thus, before the reaching of actual packets it can predict where the time slots to be given to a certain packet. By this method we can reduce the time delay of waiting of MCS level to receive the packets from UE or eNodeB. The transfer blocks will be prepared on reference to the prediction which will be done by the MCS level. The algorithm of eNodeB will be the same of the receiver section so that the retrieving of signal can be done swiftly. By this time, the UE will create multiple blocks to be matched with the prediction coming from eNodeB so that the time which will be required to produce the uplink transmission after reaching the packets can be omitted. The packet which will match the incoming package will be sent over, and the others will be discarded.

### 6.2.1 Additional Control Region for Downlink Transmission

In this Thesis, the present control region, restricted to the beginning of a subframe, is termed as primary control region. We propose that in addition to the primary control region, a secondary control region will be used throughout the rest part of the subframe. The secondary control region takes up symbols that can range from the symbol next to the primary control region up to the symbol before the last. The secondary control region will be used for quick allocation of downlink resources. Thus, the primary control region does not allocate for all resource blocks (RBs). It rather leaves RBs out in the time-frequency grid of the subframe and these RBs will be considered broken. A broken RB will contain a short resource blocks (SRB) in every symbol period as shown in Fig. 2. Thus, a SRB will have 12 resource elements (REs).

The secondary control region will allocate SRBs symbol by symbol. Like the primary control region, the secondary control region will also be comprised of strongly coded PDCCH instances and the CRC of a PDCCH instance will be scrambled by the C-RNTI of the UE. If any downlink packets for highly delay sensitive applications arrive after the primary control region is complete, the next available symbol in the secondary control region will allocate resources for them and preferably, it will allocate SRBs in its next symbol minimizing the delay. However, if the secondary control region does not find sufficient packets from highly delay sensitive applications for allocation of all SRBs in its next symbol, it can allocated from less delay sensitive applications or even from delay tolerant applications and thus, avoid wastage of SRBs.

The secondary control region itself will be allocated on SRBs and the allocation will be specified by a PDCCH instance in the primary control region. The secondary control region can be either continuously or intermittently spread over the whole subframe. The secondary control region should be allocated on frequencies that are suffering from less interference and noise. A broken RB can be shared, on different SRBs, by both the secondary control region and the downlink data allocation.

In the case of some machine type communication (MTC), especially, V2X communication, the applications are always highly delay sensitive. For such cases, a new UE category needs to be defined for the UEs. These UEs will always monitor the secondary control region for possible allocations. For the rest of the UEs in the cell, the primary control region, using PDCCH instances, will indicate which UEs need to monitor the secondary control region.

### 6.2.2 Prior Packet Generation for Uplink Transmission

The user equipment (UE) may have its data available for transmission and it cannot generate packets until it receives PDCCH signaling because the specified MCS is unknown. As shown in Fig. 1, because of the timing offset, when the UE receives PDCCH signaling on subframe n, the subframe n on PUSCH can already elapse some good extent of time, albeit less than the length of one subframe. This is because the maximum possible timing offset is 0.67 ms. Thus, the next available subframe on PUSCH is n+1. We propose that the UE generates uplink packets based on predicted MCS before the reception of PDCCH signaling. If PDCCH signaling is sent on subframe n, then it allocates resources on PUSCH on the subframe n+1. This can reduce the latency by 2.5 ms.

The eNodeB will specify which algorithm the UE should use for the prediction of MCS. The eNodeB will use the same algorithm will be made assuming the patterns of previous uplink transmission packets. Thus, before the reaching of actual packets, it can predict where the time slots to be given to a certain packet. By this method, we can reduce the time delay of waiting for MCS level to receive the packets from UE or eNodeB. The transfer blocks will be prepared with reference to the prediction which will be done by the MCS level. The algorithm of eNodeB will be the same of the receiver section so that the retrieving of the signal can be done swiftly. By this time, the UE will create multiple blocks to be matched with the prediction coming from eNodeB.

So that the time which will be required to produce the uplink transmission after reaching the packets can be omitted. The packet which will match the incoming package will be sent over, and the others will be discarded.

To evaluate the proposed scheme, in this section, we present analytical models using the approach presented in and . We assume an M/G/1 system queue.

Thus, the packet arrival follows a Poisson process and the inter-arrival times are distributed exponentially. We assume that the packet arrival rate is λ and that the service or transmission rate of the packets is μ. Thus, the mean transmission time of a packet is E[S]=1/μ. The traffic intensity can be expressed as ρ = λ/μ. We assume that the system is stable with μ > λ. The probability of packet arrival in t1 duration is

$$\int_0^{t_1} \lambda e^{-\lambda t} dt = 1 - e^{-\lambda t_1}$$

Thus, the probability of no packet arrival in duration t1 is $1 - \int_0^{t_1} \lambda e^{-\lambda t} dt = e^{-\lambda t_1}$.

We assume that the probability of packet arrival on the ith cycle is $X_i$. Since the packet arrival follows a Poisson process,$X_i$ can be expressed as

$X_i = (e^{-\lambda t_1})^{i-1}(1 - e^{-\lambda t_1})$

The symbol period is $T_C$ = 71.4 μs assuming normal cyclic prefix.

## *6.3 Calculations:*

### 6.3.1 Mathematical calculations theory:

Given

The probability of packet arrival at i-th cycle is$X_i$

Where,

$X_i = (e^{-\lambda t_1})^{i-1}(1 - e^{-\lambda t_1})$ (1)

And the expected number of packet arrival is denoted by $E[N_C]$ and it can be expressed as,

$E[N_C] = \sum_{i=1}^{14-C} i\, X_i$ (2)

Putting the value of $X_i$ from equation (1) into Equation (2) we get,

$$\mathbf{E[N_C]} = \frac{\mathbf{1 - (15 - C)e^{-\lambda Tc(14-C)} + (14 - C)e^{-\lambda Tc(15-c)}}}{\mathbf{1 - e^{-\lambda Tc}}}$$

The expected number of packet arrival is denoted by $E[N_C]$ and to get the value of expected number of packet arrival for different values of lamda the following equations which have been plotted in MATLAB to get the graphical results:

**E[Nc]={14-C-[Round E[Nc]+1}x symbol period+.5**

Here,
Symbol period=(144+2048)x3.25
=.071

### 6.3.2 MATLAB code:

```
clear all;
close all;

lamda=0:.005:1;
sym=(144+2048)/30720;
Tc=sym;

figure();
c=1;
for i=1:1:201
E_Delay(i)=(1-(15-c)*exp(-lamda(i)*Tc)^(14-c)+((14-c)*(exp(-
lamda(i)*Tc).^(15-c)))./(1-exp(-lamda(i)*Tc));
S(i)=round(E_Delay(i))+1    ;
a(i)=14-S(i);
b(i)=a(i)*sym+.5;
end
plot(lamda,b);
xlabel('For C=1, E Delay vs lamda');
hold off


c=2;
figure();
for i=1:1:101
E_Delay(i)=(1-(15-c)*exp(-lamda(i)*Tc)^(14-c)+((14-c)*(exp(-
lamda(i)*Tc).^(15-c)))./(1-exp(-lamda(i)*Tc));
S(i)=round(E_Delay(i))+1    ;
a(i)=14-S(i);
b(i)=a(i)*sym+.5;
end
plot(lamda,b);
xlabel('For C=2, E Delay vs lamda');
hold off

c=3;
figure();
for i=1:1:101
E_Delay(i)=(1-(15-c)*exp(-lamda(i)*Tc)^(14-c)+((14-c)*(exp(-
lamda(i)*Tc).^(15-c)))./(1-exp(-lamda(i)*Tc));
S(i)=round(E_Delay(i))+1    ;
a(i)=14-S(i);
b(i)=a(i)*sym+.5;
end
plot(lamda,b);
xlabel('For C=3, E Delay vs lamda');
hold off
```

### 6.3.3 Experimentally found results:

Table: 6. 1For C=1

| Lamda | E[Nc] |
|---|---|
| **0.005** | 1.356 |
| **0.08** | 1.356 |
| **0.085** | 1.285 |
| **0.27** | 1.285 |
| **0.275** | 1.214 |
| **0.525** | 1.214 |
| **0.53** | 1.142 |
| **0.92** | 1.142 |
| **0.925** | 1.071 |
| **1.00** | 1.071 |

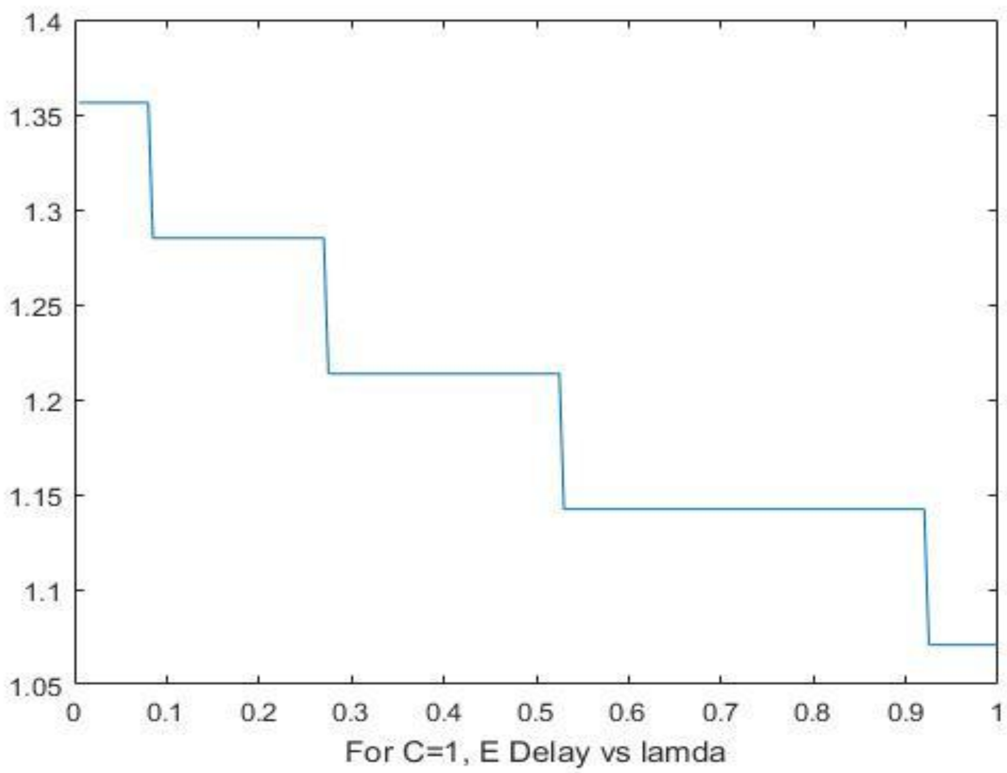


Figure 6-1 For C=1

Table:6.2 For C=2

| Lamda | E[Nc] |
|---|---|
| **0.005** | 1.428 |
| **0.09** | 1.428 |
| **0.095** | 1.356 |
| **0.32** | 1.356 |
| **0.325** | 1.285 |
| **0.525** | 1.285 |
| **0.530** | 1.214 |
| **0.92** | 1.214 |
| **0.925** | 1.142 |
| **1.00** | 1.142 |

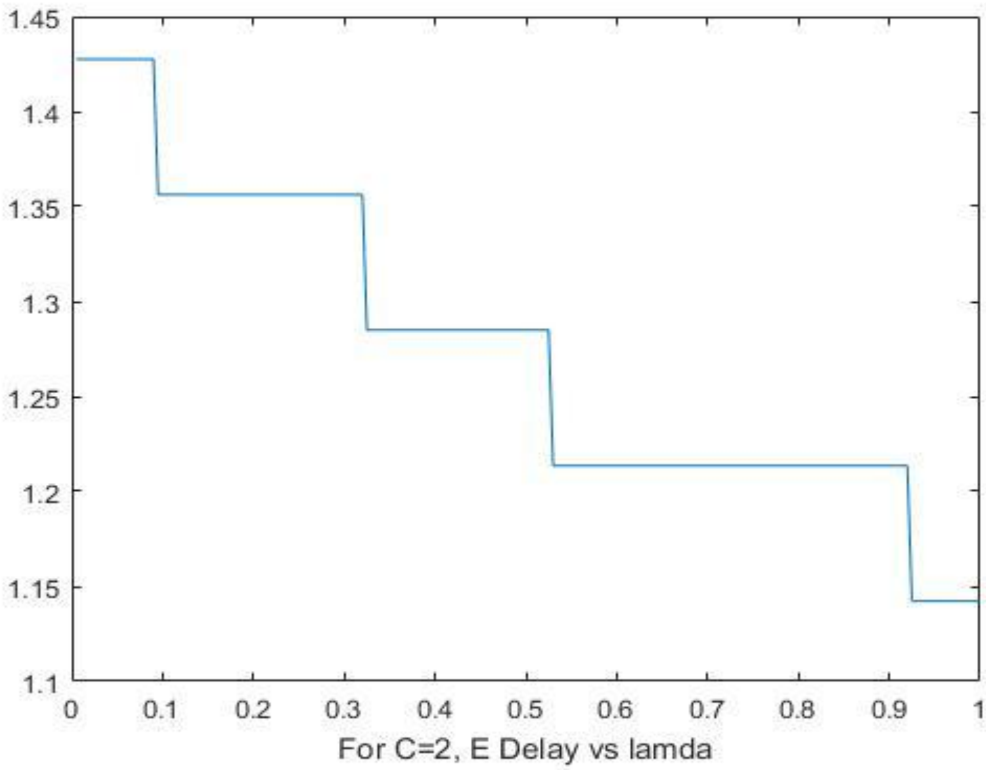


Figure 6-2For C=2

Table: 6. 3 For C=3

| Lamda | E[Nc] |
|---|---|
| | |
| **0.005** | 1.428 |
| **0.11** | 1.428 |
| **0.115** | 1.356 |
| **0.385** | 1.356 |
| **0.390** | 1.285 |
| **0.525** | 1.285 |
| **0.53** | 1.214 |
| **0.92** | 1.214 |
| **1.925** | 1.142 |
| **1.00** | 1.142 |

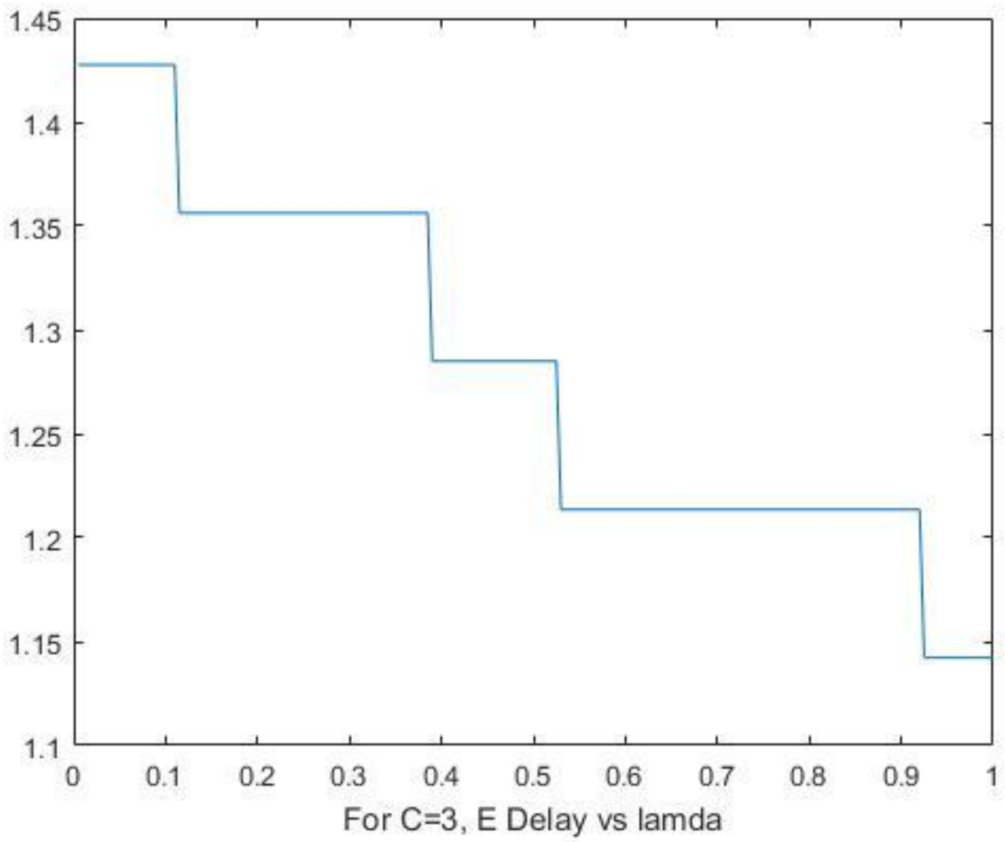


Figure 6-3For C=3

### 6.3.4 Previous Delay Calculations

Table6. 4 Previous Delay Table

| Delay Components | Delay Value |
|---|---|
| Transmission time uplink+Downlink | 2 ms |
| Buffering time ( 0.5 * transmission time | 2*.5=1 ms |
| Retransmission | 2*.1*8=1.6 ms |
| Uplink scheduling request | .4625*5=2.31 ms |
| Uplink scheduling grant | 0 ms |
| UE delay estimated | 4 ms |
| EnodeB dealy estimated | 4 ms |
| Core Network | 1 ms |
| Total delay with pre allocated resource | 13.6 ms |
| Total delay with scheduling | 15.91 ms |

### 6.3.5 Improved delay calculations:

Table6. 5 Improved Delay Table:

| **Delay Components** | **Delay Value** |
|---|---|
| Transmission Time Uplink+ Downlink | 2 ms |
| Buffering time (0.5 x transmission time) | 2*.5*1=1 ms |
| Retransmission (10%) | 2*1.8*1=1.6 ms |
| Uplink scheduling request | .465*5=2.31 ms |
| Uplink scheduling grant | 1 ms |
| UE delay estimated | 4 ms |
| eNodeB delay estimated | 4ms |
| Core network | 1 ms |
| Total delay with pre allocated resource | 13.6 ms |
| Total delay with scheduling | 15.91 ms |

# Chapter 7

# Conclusion

5G radio access technology will be a key component of the Networked Society. It will address high traffic growth and increasing demand for high-bandwidth connectivity. It will also support massive numbers of connected devices and meet the real-time, high-reliability communication needs of mission-critical applications. 5G will provide wireless connectivity for a wide range of new applications and use cases, including wearables, smart homes, traffic safety/control, critical infrastructure, industry processes and very-high-speed media delivery. As a result, it will also accelerate the development of the Internet of Things. The overall aim of 5G is to provide ubiquitous connectivity for any kind of device and any kind of application that may benefit from being connected. 5G networks will not be based on one specific radio-access technology. Rather, 5G is a portfolio of access and connectivity solutions addressing the demands and requirements of mobile communication beyond 2020.

We analyzed the latency of call setup from idle for a legacy WCDMA network where signaling is carried on dedicated channels. We showed that a significant (~1.2 sec) gain in latency of call setup from idle can be achieved by carrying signaling on HSPA channels. We showed our two proposals in this regard. Proposal 1: Dual Control Plane: The two control planes PDCCH and PUCCH may allocate resources to be used by the UE and eNodeB. The TTI is thus predefined by symbols predicted by the control channel. Present control region may indicate which PDCCH format should be used. Present control region will indicate which symbol in Control region 2 should use which RNTI. Control region 1 will allocate resources for some delay tolerant apps and leave some delay sensitive apps to be allocated by control region 2. And Proposal 2:Prediction of uplink transmission: Packets will be made assuming the patterns of previous uplink transmission packets. Thus, before the reaching of actual packets it can predict where the time slots to be given to a certain packet. By this method we can reduce the time delay of waiting of MCS level to receive the packets from UE or eNodeB. The transfer blocks will be prepared on reference to the prediction which will be done by the MCS level .We showed that LTE further reduces the latency of call setup from idle latency by ~360 ms compared to an HSPA system where signaling is carried on HSPA channels. One of the key reasons that contribute to this gain for LTE is that the Radio Bearer Setup message is assumed to be a synchronized message for HSPA, whereas the corresponding message in LTE is unsynchronized.

Unsynchronized Radio Bearer Setup is supported by the HSPA specification, but we did not consider it in this paper since it is not commonly deployed. We also compared the latencies of call setup from connected between HSPA and LTE systems and showed them to be similar, if the Enhanced CELL_PCH state, introduced in Release 7 of the HSPA specification, is considered. If starting from legacy CELL_PCH, HSPA latency is 69-75 ms more than that of LTE. We expect that UEs will typically camp in connected camped state (i.e., CELL_PCH for HSPA and connected with long DRX cycle for LTE), and thus latency of call setup from connected is likely to drive the user's perception of the responsiveness of the system when the UE is camped. Finally, we compared the U-plane latency (or PING request-response latency) of HSPA and LTE systems. We showed that, with similar H-ARQ operating points and similar processing latencies of network nodes, the U-plane latencies for HSPA and LTE are similar.

The specification of 5G will include the development of a new flexible air interface, NX, which will be directed to extreme mobile broadband deployments. NX will also target high-bandwidth and high-traffic-usage scenarios, as well as new scenarios that involve mission-critical and real time communications with extreme requirements in terms of latency and reliability. In parallel, the development of Narrow-Band IoT (NB-IoT) in 3GPP is expected to support massive machine connectivity in wide area applications. NB-IoT will most likely be deployed in bands below 2GHz and will provide high capacity and deep coverage for enormous numbers of connected devices.

Ensuring interoperability with past generations of mobile communications has been a key principle of the ICT industry since the development of GSM and later wireless technologies within the 3GPP family of standards. In a similar manner, LTE will evolve in a way that recognizes its role in providing excellent coverage for mobile users, and 5G networks will incorporate LTE access (based on Orthogonal Frequency Division Multiplexing (OFDM)) along with new air interfaces in a transparent manner toward both the service layer and users. Around 2020, much of the available wireless coverage will continue to be provided by LTE, and it is important that operators with deployed 4G networks have the opportunity to transition some – or all – of their spectrum to newer wireless access technologies. For operators with limited spectrum resources, the possibility of introducing 5G capabilities in an interoperable way – thereby allowing legacy devices to continue to be served on a compatible carrier – is highly beneficial and, in some cases, even vital.

V2X communications is a critical component of the connected car of the future. Cellular V2X brings improvements to DSRC for safety use cases and beyond. Cellular V2X vehicles will have superior capabilities, both for the direct mode and the network mode. Cellular V2X is opening up new opportunities and business models for mobile operators, resulting in advanced services provided for transportation users. In light of LTE's evolution toward 5G and V2X, it is becoming evident that vehicles connected to the cellular network will be given superior capabilities and possibilities, especially when considering V2N architecture approach. The network can reach vehicles within a large region, allowing for increased range, and for predictive and early measures towards safer traffic flow. With this in mind, the cellular ecosystem stakeholders should engage in early efforts to assess the forthcoming capacity and coverage needs of connected vehicles. Together, as partners with the vehicle industry.

5G is the next step in the evolution of mobile communication and will be a key component of the Networked Society. In particular, 5G will accelerate the development of the Internet of Things. To enable connectivity for a wide range of applications and use cases, the capabilities of 5G wireless access must extend far beyond those of previous generations of mobile communications. These capabilities include very high achievable data rates, very low latency and ultra-high reliability. Furthermore, 5G wireless access needs to support a massive increase in traffic in an affordable and sustainable way, implying a need for a dramatic reduction in the cost and energy consumption per delivered bit. 5G wireless access will be realized by the evolution of LTE for existing spectrum in combination with new radio access technologies that primarily target new spectrum. Key technology components of 5G wireless access include access/backhaul integration, device-to-device communication, flexible duplex, flexible spectrum usage, multi-antenna transmission, ultra-lean design, and user/control separation.

# References

[1] Latency Reduction Technique for Next Generation Cellular Network Jeongho Yeo, Jinyoung Oh, Younsun Kim, Juho Lee, STE Team, DMC R&D Center, Samsung Electronics, Seongchon-gil 34, Seocho-gu, Seoul, 06765, South Korea.

[2] 3G Americas – "Transition to 4G, 3GPP Broadband Evolution to IMT-Advanced"- http://www.ngmn.org/fileadmin/user_upload/News/Partner_News/3G_Americas_Rysavy Research_HSPA-LTE_Advanced_FINAL.pdfqualcomm-research-latency-in-hspa-data-networks.pdf

[3] 3GPP TS 25.214, "Physical Layer Procedures (FDD)".

[4] 3GPP TS 36.214, "Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer; Measurements". "3GPP TS 25.331, "Radio Resource Control (FDD)".

[5] 3GPP TS 36.331, "Evolved Universal Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification"

[6] Ericsson White paper, Uen 284 23-3204 Rev C | April 2016," 5G radio access".

[7] Avilable:https://www.ericsson.com/assets/local/publications/white-papers/wp-5g.pdf [ Accesed 2017].

[8] GSMA intelligence, "Understanding 5G: Perspectives on future technological advancements in mobile". December 2014.

[9] Analysis and Simulation of LTE Downlink and Uplink Transceiver
Temitope O. Takpor and Francis E. Idachaba, Members, IAENG.

[10] Predictability of Uplink Channels From Downlink Channels in Frequency Division Duplex system: A power control application.
D. Mihai Ionescu, Senior Member, IEEE, and Adrian Boariu, Member, IEEE.

[11] Uplink Control Channel Design for 5G Ultra-LowLatency Communication
Shuqiang Xia, Xianghui Han, Xiao Yan, Zhisong Zuo, Feng Bi

[12] Conversion from Uplink to Downlink Spatio-Temporal Correlation with Cubic Splines
Markus Jordan, Andrei Dimofte, Xitao Gong, and Gerd Ascheid

[13] Achieving low latency and energy consumption by 5G TDD mode optimization
Eeva Lähetkangas (1), Kari Pajukoski (1), Jaakko Vihriälä (1), Gilberto Berardinelli (2), Mads Lauridsen (2), Esa Tiirola (1), Preben Mogensen (2,3)